盘古大模型

API 参考

文档版本 01

发布日期 2025-11-03

华为云计算技术有限公司

商标声明


和其他华为商标均为华为技术有限公司的商标。
本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为云计算技术有限公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为云计算技术有限公司对本文档内容不做任何明示或暗示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

华为云计算技术有限公司

地址： 贵州省贵安新区黔中大道交兴功路华为云数据中心 邮编：550029

网址： https://www.huaweicloud.com/

目 录

1 使用前必读

1.1 概述

ModelArts Studio大模型开发平台支持纳管三方大模型，模型在平台部署后，可以通过API调用推理接口。

表 1-1 API 清单

类别	模型	API	功能
模型推理接口	三方NLP大模型	三方NLP大模型	DeepSeek API是基于DeepSeek大模型推出的接口服务，它支持多场景文本交互，能够快速生成高质量对话、文案、故事等内容，可用于文本摘要、智能问答、内容创作等场景。
	Qwen三方VL大模型	Qwen三方VL大模型	Qwen2.5-VL系列模型，具备图像识别、精准视觉定位、文字识别和理解、文档解析、视频理解等能力。
数据工程接口	-	查询数据血缘	客户通过obs导入原始数据集，可基于该obs路径查询所有基于该路径创建的原始数据集及后续的血缘信息。
	-	数据集彻底删除	只针对从obs上传的数据，在删除数据集的时候要关联删除OBS下对应的原始数据，客户认为原始数据应该在客户侧大数据中心长期归档，不应该在OBS长期保留。
Agent应用接口	-	调用应用	通过调用创建好的应用API，输入问题，将得到应用执行的结果。
	-	调用工作流	通过调用创建好的工作流API，输入问题，将得到工作流执行的结果。

类别	模型	API	功能
Token计算器	-	Token计算器	为了帮助用户更好地管理和优化Token消耗，平台提供了Token计算器工具。Token计算器可以帮助用户在模型推理前评估文本的Token数量，提供费用预估，并优化数据预处理策略。

说明

用户在部署服务的过程中，建议开启“安全护栏”功能，以保证内容的安全性。

1.2 调用说明

盘古大模型提供了REST（Representational State Transfer）风格的API，支持您通过HTTPS请求调用，调用方法请参见**如何调用REST API**。

调用API时，需要用户网络可以访问公网。

1.3 请求 URI

服务的请求URI即API服务的终端地址，通过该地址与API进行通信和交互。

URI获取步骤如下：

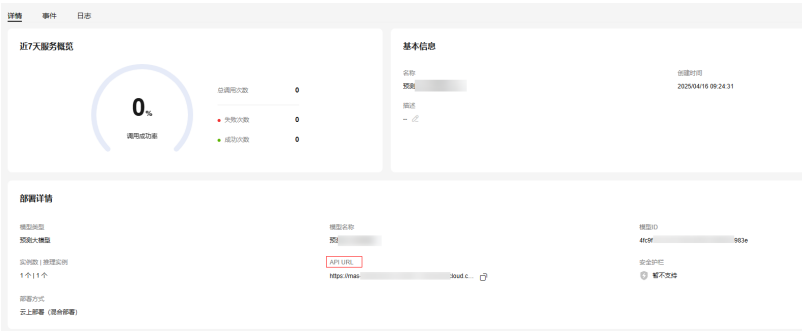
步骤1 登录ModelArts Studio大模型开发平台。

步骤2 进入所需工作空间。

步骤3 获取请求URI。

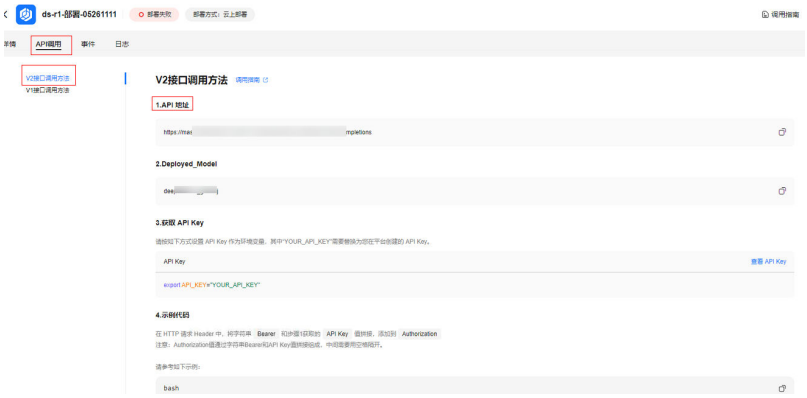
- 获取模型请求URI。
  - 若调用部署后的模型，可在左侧导航栏中选择“模型开发 > 模型部署”，在“我的服务”页签，模型部署列表单击模型名称，在“详情”页签中，可获取模型的请求URI。

图 1-1 部署后的模型调用路径



若调用的是用户自己部署的NLP推理服务，可在“API调用”页签中，可获取V1接口URL或者V2接口URI。

图 1-2 NLP 服务的调用路径



- 若调用预置模型，可在左侧导航栏中选择“模型开发 > 模型部署”，在“预置服务”页签，模型列表单击“调用路径”，获取该模型的请求URI。

图 1-3 预置模型的调用路径

- 获取Agent应用请求URI。
  - 单击左侧导航栏“Agent开发”，进入“工作台 > 应用”页面，选择需要部署的应用，单击“ ··· > 调用路径”。
  - 在“调用路径”页面可获取Agent应用请求URI。

图 1-4 调用路径

----结束

1.4 基本概念

- 账号

  用户注册华为云时的账号，账号对其所拥有的资源及云服务具有完全的访问权限，可以重置用户密码、分配用户权限等。由于账号是付费主体，为了确保账号安全，建议您不要直接使用账号进行日常管理工作，而是创建用户并使用他们进行日常管理工作。

- 用户

  由账号在IAM中创建的用户，是云服务的使用人员，具有身份凭证（密码和访问密钥）。

  在**我的凭证**下，您可以查看账号ID和用户ID。通常在调用API的鉴权过程中，您需要用到账号、用户和密码等信息。

- 区域（Region）

  从地理位置和网络时延维度划分，同一个Region内共享弹性计算、块存储、对象存储、VPC网络、弹性公网IP、镜像等公共服务。Region分为通用Region和专属Region，通用Region指面向公共租户提供通用云服务的Region；专属Region指只承载同一类业务或只面向特定租户提供业务服务的专用Region。

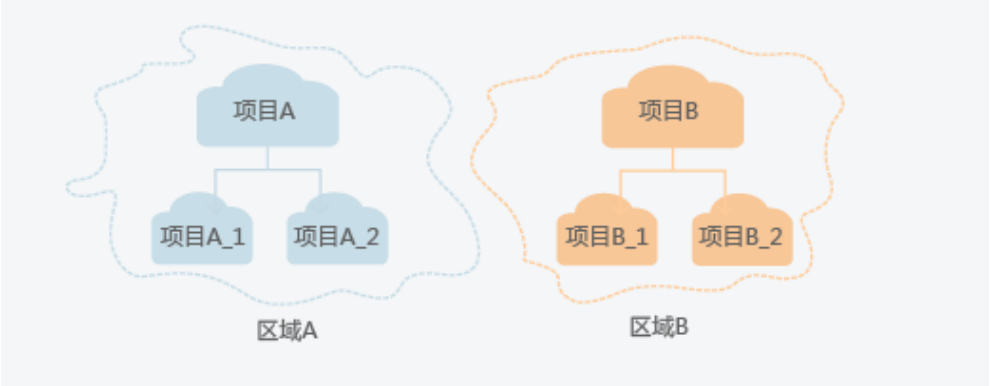
- 可用区（AZ，Availability Zone）

  一个AZ是一个或多个物理数据中心的集合，有独立的风火水电，AZ内逻辑上再将计算、网络、存储等资源划分成多个集群。一个Region中的多个AZ间通过高速光纤相连，以满足用户跨AZ构建高可用性系统的需求。

- 项目

  华为云的区域默认对应一个项目，这个项目由系统预置，用来隔离物理区域间的资源（计算资源、存储资源和网络资源），以默认项目为单位进行授权，用户可以访问您账号中该区域的所有资源。如果您希望进行更加精细的权限控制，可以在区域默认的项目中创建子项目，并在子项目中购买资源，然后以子项目为单位进行授权，使得用户仅能访问特定子项目中资源，使得资源的权限控制更加精确。

图 1-5 项目隔离模型



2 如何调用 REST API

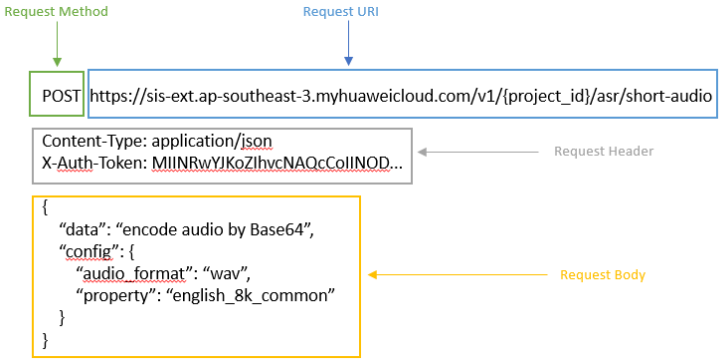
2.1 构造请求

本节介绍REST API请求的组成，并以调用服务的**获取用户Token**接口说明如何调用API。

请求示例如**图2-1**，一个请求主要由**请求URI**、**请求方法**、**请求消息头**和**请求消息体**组成。

图 2-1 请求示例图

图 2-2 请求示例图



请求 URI

请求URI由如下部分组成：

{URI-scheme}://{endpoint}/{resource-path}?{query-string}

表 2-1 请求 URI

参数	说明
URI-scheme	传输请求的协议，当前所有API均采用**HTTPS**协议。
endpoint	承载REST服务端点的服务器域名或IP。
resource-path	资源路径，即API访问路径。从具体API的URI模块获取。
query-string	查询参数，可选，查询参数前面需要带一个“？”，形式为“参数名=参数取值”。

获取请求URI的步骤详见**请求URI**，示例如下：

```
https://{endpoint}/v1/{project_id}/deployments/{deployment_id}/chat/completions
```

请求方法

HTTP请求方法，表示服务正在请求操作类型，包括：

- **GET**：请求服务器返回指定资源。
- **PUT**：请求服务器更新指定资源。
- **POST**：请求服务器新增资源或执行特殊操作。
- **DELETE**：请求服务器删除指定资源，如删除对象等。
- **HEAD**：请求服务器资源头部。
- **PATCH**：请求服务器更新资源的部分内容。当资源不存在的时候，PATCH可能会去创建一个新的资源。

在接口的URI部分，请求方法为“POST”，例如：

```
POST https://{endpoint}/v1/{project_id}/deployments/{deployment_id}/chat/completions
```

请求消息头

附加请求头字段，如指定的URI和HTTP方法所要求的字段。例如，定义消息体类型的请求头“Content-Type”，请求鉴权信息等。

以下公共消息头需要添加到请求中。

- **Content-Type**：消息体的类型（格式），必选，默认取值为“application/json”。
- **X-Auth-Token**：用户Token，可选，当使用Token方式认证时，必须填充该字段。用户Token请参考**认证鉴权**中的“Token认证”。

  说明

  公有云API同时支持使用AK/SK认证，AK/SK认证是使用SDK对请求进行签名，签名过程会自动往请求中添加Authorization（签名认证信息）和X-Sdk-Date（请求发送的时间）请求头。AK/SK认证的详细说明请参见：**AK/SK**。

添加消息头后的请求如下所示：

```
POST https://{endpoint}/v1/{project_id}/deployments/{deployment_id}/chat/completions
Content-Type: application/json
X-Auth-Token: MIINRwYJKoZIhvcNAQcCoIINOD...
```

请求消息体

请求消息体通常以结构化格式发出，与请求消息头中Content-Type对应，传递除请求消息头之外的内容。若请求消息体中参数支持中文，则中文字符必须为UTF-8编码。

每个接口的请求消息体内容不同，也并不是每个接口都需要有请求消息体（或者说消息体为空），GET、DELETE操作类型的接口就不需要消息体，消息体具体内容需要根据具体接口而定。

将消息体加入后的请求如下所示，详细参数解释可参考文档API章节。

```
POST https://{endpoint}/v1/{project_id}/deployments/{deployment_id}/chat/completions
Content-Type: application/json
X-Auth-Token: MIINRwYJKoZIhvcNAQcCoIINOD...
{
   "messages": [
    {
        "content": "介绍下长江，以及长江中典型的鱼类"
    }
    ],
   "temperature": 0.9,
   "max_tokens": 600
}
```

综上，您可以使用**curl**、**Postman**或直接编写代码等方式发送请求调用API。对于接口，您可以从响应消息部分看到返回参数及参数说明。

2.2 认证鉴权

调用接口有如下认证方式，您可以选择其中一种进行认证鉴权。

- Token认证：通过Token认证调用请求。
- API Key认证：当用户部署的模型服务期望开放给其他用户调用时，使用原有的Token认证需要进行动态认证鉴权和凭证管理，操作繁杂。此时可使用API Key认证。该方式不仅相比Token认证更简便，还与业界主流模型调用规范保持一致。

Token 认证

Token在计算机系统中代表令牌（临时）的意思，拥有Token就代表拥有某种权限。Token认证就是在调用API的时候将Token加到请求消息头，从而通过身份认证，获得操作API的权限。

说明

- Token的有效期为24小时，需要使用一个Token鉴权时，可以先缓存，避免频繁调用。
- 如果您的华为云账号已升级为华为账号，将不支持获取账号Token。建议为您自己创建一个IAM用户，获取IAM用户的Token。

获取Token方法：

Token可通过调用“获取Token”接口获取，接口调用示例如下：

- 伪码

```
POST https://iam.ap-southeast-1.myhuaweicloud.com/v3/auth/tokens //以获取中国-香港区域Token为例
Content-Type: application/json
{
    "auth": {
        "identity": {
            "methods": [
                "password"
            ],
            "password": {
                "user": {
                    "name": "username", //IAM用户名
                    "password": "********", //IAM用户密码
                    "domain": {
                        "name": "domainname" //账号名
                    }
                }
            }
        },
        "scope": {
            "project": {
                "name": "ap-southeast-1" //盘古大模型当前部署在中国-香港区域，取值为ap-southeast-1
            }
        }
    }
}
```

- Python

```
import requests
import json

url = "https://iam.ap-southeast-1.myhuaweicloud.com/v3/auth/tokens"
payload = json.dumps({
  "auth": {
    "identity": {
      "methods": [
        "password"
      ],
      "password": {
        "user": {
          "name": "username",
          "password": "********",
          "domain": {
            "name": "domainname"
          }
        }
      }
    },
    "scope": {
      "project": {
        "name": "projectname"
```

```
    }
   }
  }
})
headers = {
  'Content-Type': 'application/json'
}

response = requests.request("POST", url, headers=headers, data=payload)

print(response.headers["X-Subject-Token"])
```

获取Token步骤：

本示例中，通过使用Postman软件获取Token。

1. 登录“我的凭证 > API凭证”页面，获取user name、domain name、project id。

   由于盘古大模型当前部署在中国-香港区域，需要获取与中国-香港区域对应的project id。

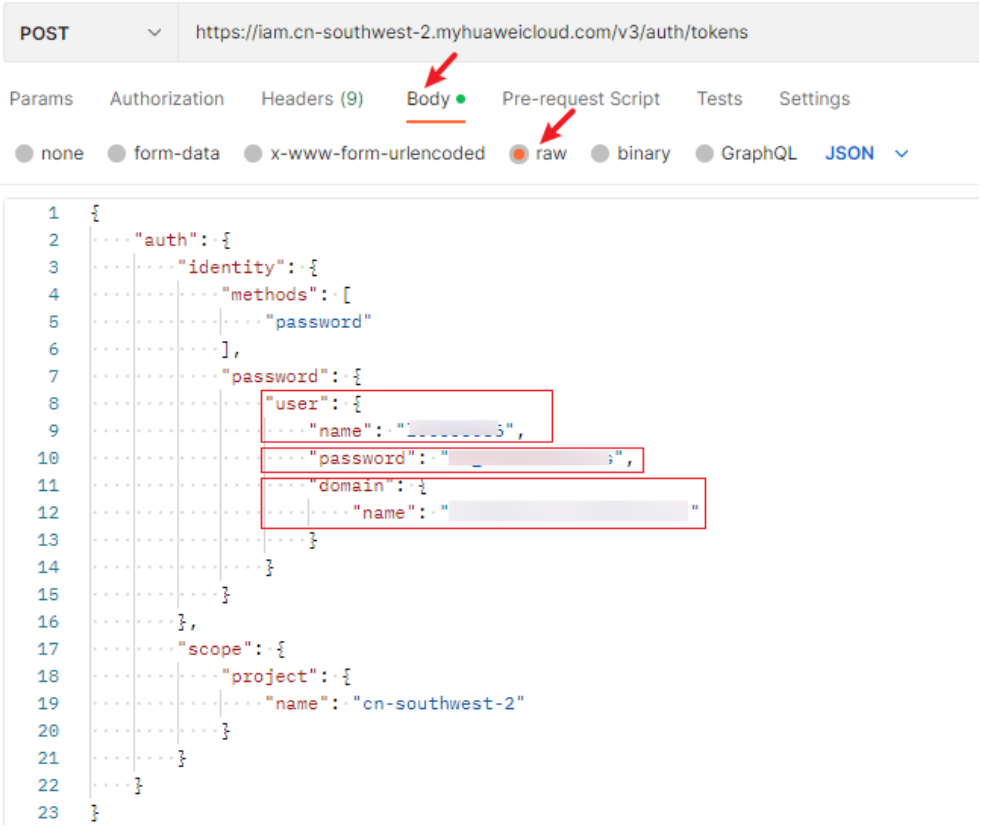
   图 2-3 获取 user name、domain name、project id

2. 打开Postman，新建一个POST请求，并输入中国-香港区域的“获取Token”接口。并填写请求Header参数。
   - 接口地址为：https://iam.ap-southeast-1.myhuaweicloud.com/v3/auth/tokens
   - 请求Header参数名为Content-Type，参数值为application/json
3. 填写“获取token”接口的请求体。在Postman中选择“Body > raw”选项，参考图2-4复制并填入以下代码，并填写user name、domain name、password。

```
{
    "auth": {
        "identity": {
            "methods": [
                "password"
            ],
            "password": {
                "user": {
                    "name": "username", //IAM用户名
                    "password": "********", //华为云账号密码
                    "domain": {
                        "name": "domainname" //账号名
                    }
                }
            }
        },
        "scope": {
            "project": {
                "name": "ap-southeast-1" //盘古大模型当前部署在“中国-香港”区域，取值为ap-southeast-1
            }
        }
    }
}
```

图 2-4 填写请求 Body



4. 单击Postman界面“Send”按钮，发送请求。当接口返回状态为201时，表示Token接口调用成功，此时单击“Headers”选项，找到并复制“X-Subject-Token”参数对应的值，该值即为需要获取的Token。

API Key 认证

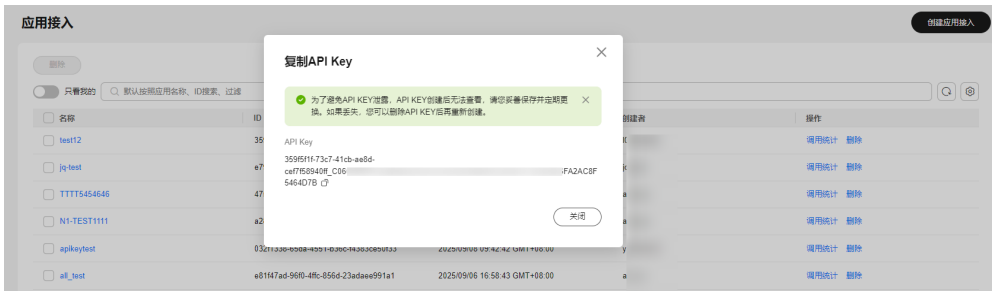
当用户部署的API服务期望开放给其他用户调用时，原有**Token认证**无法支持，可使用API Key认证的鉴权方式进行调用请求。

API Key认证指调用API时，在HTTP请求头部消息增加一个参数X-Apig-AppCode（参数值为“API Key”值），而不需要对请求内容签名。

使用该鉴权方式前，请确保有已部署的大模型。

获取API Key步骤如下：

1. 登录ModelArts Studio平台，进入所需空间。
2. 在左侧导航栏中选择“应用接入”，单击界面右上角“创建应用接入”。
3. 在“关联服务”中，选择“全部服务”，或者在“指定服务”中选择已部署好的大模型服务，单击“确定”。
4. 在弹窗中复制API Key的值。注意：API Key仅可在此弹窗中复制一次，请妥善保存。关闭弹窗后，如果忘记API Key值，请重新创建。



2.3 返回结果

状态码

请求发送以后，您会收到响应，包含状态码、响应消息头和消息体。

状态码是一组从1xx到5xx的数字代码，状态码表示了请求响应的状态，完整的状态码列表请参见**状态码**。

对于Pangu服务接口，如果调用后返回状态码为“200”，则表示请求成功。

响应消息头

对应请求消息头，响应同样也有消息头，如“Content-Type”。

响应消息体

响应消息体通常以结构化格式返回，与响应消息头中Content-Type对应，传递除响应消息头之外的内容。

接口调用成功后将返回如下响应体。


```
{
   "id": "180f5745-4ee4-42a9-9869-23f829654bb7",
   "created": 1724915285,
   "choices": [
      {
         "index": 0,
         "text": "很久很久以前，在一个遥远的王国里，有一个美丽的国度叫做艾尔多利亚。艾尔多利亚是一个充满绿意盎然的森林、闪闪发光的河流和雄伟山脉的王国。这个王国由一位英明公正的国王阿尔里克和他的王后伊索尔德统治。他们有一个儿子，王子艾登，他勇敢善良，还有一个女儿，公主伊拉拉，她美丽聪慧。
艾登王子和伊拉拉公主关系非常亲密。他们一起在花园里玩耍，探索森林，学习关于他们的王国。随着年龄的增长，他们的关系更加深厚，他们意识到彼此已经坠入爱河。
然而，他们的爱情并非没有挑战。阿尔里克国王和伊索尔德王后为他们的孩子制定了其他计划。他们希望艾登王子与邻国的公主结婚以结成联盟，希望伊拉拉公主与遥远国度的王子结婚以增强王国的力量。
艾登王子和伊拉拉公主心碎了。他们知道必须做些什么才能在一起。一天晚上，他们决定逃离宫殿，私奔。他们在黑暗中悄悄离开宫殿，带着行李踏上了旅程。
他们穿越森林和山脉，一路上遇到了许多危险。但他们对彼此的爱给了他们克服每一个障碍的力量。最终，他们到达了艾尔多利亚边缘的一个小村庄，决定在那里定居。
他们在村庄里过着简单的生活，辛勤工作，深深相爱。他们很幸福，但他们知道不能永远隐藏下去。有一天，阿尔里克国王和伊索尔德王后发现他们逃跑的消息，来到村庄接他们回去。
艾登王子和伊拉拉公主恳求父母让他们在一起。他们讲述了彼此的爱情，以及他们无法想象没有对方的生活。阿尔里克国王和伊索尔德王后被他们的爱情感动，决定让他们在一起。
他们回到宫殿，向王国宣布了他们的决定。艾尔多利亚的人民欢欣鼓舞，庆祝艾登王子和伊拉拉公主的结合。他们的爱情故事成为了艾尔多利亚的传奇。
艾登王子和伊拉拉公主一起统治艾尔多利亚，为王国带来了和平与繁荣。他们的爱情故事激励了许多人，他们的王国在他们统治下繁荣昌盛。他们幸福地生活在一起，证明了真爱可以克服任何障碍。",
         "ppl": 1.77809815678146e-36
      }
   ],
```



```
    "usage": {
        "completion_tokens": 365,
        "prompt_tokens": 9,
        "total_tokens": 374
    }
}
```

当接口调用出错时，会返回错误码及错误信息说明。

```
token有效期为24小时，下面的报错表示token过期。
{
    "error_msg": "Incorrect IAM authentication information: token expires,
expires_at:2023-06-29T02:16:41.581000Z",
    "error_code": "APIG.0301",
    "request_id": "469967f55e6b225xxx"
}
```

其中，error_code表示错误码，error_msg表示错误描述信息。

3 API

3.1 模型推理接口

3.1.1 三方大模型

3.1.1.1 三方 NLP 大模型

功能介绍

三方NLP大模型API是基于DeepSeek和通义千问大模型推出的接口服务，它支持多场景文本交互，能够快速生成高质量对话、文案、故事等内容，可用于文本摘要、智能问答、内容创作等场景。

URI

NLP推理服务支持使用盘古推理接口（V1推理接口）调用，也支持使用业界通用的OpenAi格式接口（V2推理接口）调用。

V1接口、V2接口的鉴权方式不同，请求体和返回体略有差异。

表 3-1 NLP 服务推理接口

API分类	API访问路径（URI）
V1推理接口	POST /v1/{project_id}/deployments/{deployment_id}/chat/completions
V2推理接口	POST /api/v2/chat/completions

表 3-2 V1 推理接口路径参数

参数	是否必选	参数类型	描述
project_id	是	String	**参数解释：** 项目ID，获取方法请参见**获取项目ID**。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
deployment_id	是	String	**参数解释：** 模型的部署ID，获取方法请参见**获取模型部署ID**。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

请求参数

V1、V2推理接口的鉴权方式不同，请求参数与响应参数也有不同，说明如下：

Header参数

1. V1接口支持Token鉴权方式，也支持API Key鉴权方式。两种鉴权方式请求Header参数说明如下：
   - 使用**Token认证**方式的请求Header参数见**表3-3**。

表 3-3 请求 Header 参数（Token 认证）

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	**参数解释：** 用户Token。 用于获取操作API的权限。如**Token认证**中响应消息头中X-Subject-Token的值即为Token。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
Content-Type	是	String	**参数解释：** 发送的实体的MIME类型。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** application/json

- 使用**API Key认证**方式的请求Header参数见**表 请求Header参数（API Key认证）**。

表 3-4 请求 Header 参数（API Key 认证）

参数	是否必选	参数类型	描述
X-Apig-AppCode	是	String	**参数解释：** API Key值。 用于获取操作API的权限。**API Key认证**响应消息头中X-Apig-AppCode的值即为API Key。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

参数	是否必选	参数类型	描述
Content-Type	是	String	**参数解释：** 发送的实体的MIME类型。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** application/json

2. V2接口只支持API Key鉴权方式。请求Header参数见表3-5。

表 3-5 V2 接口请求 Header 参数（ OpenAI 格式的 API Key 认证 ）

参数	是否必选	参数类型	描述
Authorization	是	String	**参数解释：** 用户创建应用接入获取的API Key，拼接“Bearer ”后的字符串。示例：Bearer d59******9C3 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
Content-Type	是	String	**参数解释：** 发送的实体的MIME类型。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** application/json

请求Body参数

V1、V2推理接口请求Body参数一致，如表3-6。

表 3-6 请求 Body 参数

参数	是否必选	参数类型	描述
messages	是	Array of **ChatCompletionMessageParam** objects	**参数解释：** 多轮对话问答对，包含两个属性：role和content。 • role表示对话的角色，取值是system或user。 如果需要模型以某个人设形象回答问题，可以将role参数设置为system。不使用人设时，可设置为user。在一次会话请求中，人设只需要设置一次。 • content表示对话的内容，可以是任意文本。 messages参数可以帮助模型根据对话的上下文生成合适的回复。 **约束限制：** 数组长度：**1 - 20** **取值范围：** 不涉及 **默认取值：** 不涉及
model	是	String	**参数解释：** 使用的模型ID，根据所部署的模型填写，填写DeepSeek-R1或DeepSeek-V3。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

参数	是否必选	参数类型	描述
stream	否	boolean	**参数解释：** 流式开关。流式输出协议为SSE(Server-Sent Events ）协议。 如果开启流式，请赋值true。开启流式开关后，API会在生成文本的过程中，实时地将生成的文本发送给客户端，而不是等到生成完成后一次性将所有文本发送给客户端。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** **false**
temperature	否	Float	**参数解释：** 用于控制生成文本的多样性和创造力。 控制采样随机性的浮点数。一般来说，temperature越低，适合完成确定性的任务。temperature越高，如0.9，适合完成创造性的任务。值为 0 意味着贪婪采样。当取值超过1，会大概率出现效果不可用问题。 temperature参数可以影响语言模型输出的质量和多样性，但也不是唯一的因素。还有其他一些参数，如top_p参数也可以用来调整语言模型的行为和偏好，但不建议同时更改这两temperature和top_p。 **约束限制：** 不涉及 **取值范围：** (0, 1] **默认取值：** **1.0**

参数	是否必选	参数类型	描述
top_p	否	Float	**参数解释：** 核采样参数。作为调节采样温度的替代方案，模型会考虑前top_p概率的token的结果。0.1就意味着只有包括在最高10%概率中的token会被考虑。 建议修改这个值或者更改temperature，但不建议同时对两者进行修改。 **说明** token是指模型处理和生成文本的基本单位。token可以是词或者字符的片段。模型的输入和输出的文本都会被转换成token，然后根据模型的概率分布进行采样或者计算。 **约束限制：** 不涉及 **取值范围：** **（ 0.0, 1.0]** **默认取值：** **0.8**
max_tokens	否	Integer	**参数解释：** 生成文本的最大输出token数量。 **约束限制：** 输入的文本加上生成的文本总量不能超过模型所能处理的最大长度。 **取值范围：** ● 最小值：**1** ● 最大值：**8192** **默认取值：** **4096**

参数	是否必选	参数类型	描述
presence_penalty	否	Float	**参数解释：** 用于调整模型对新Token的处理方式。即如果一个Token已经在之前的文本中出现过，那么模型在生成这个Token时会受到一定的惩罚。当presence_penalty的值为正数时，模型会更倾向于生成新的、未出现过的Token，即模型会更倾向于谈论新的话题。 **约束限制：** 不涉及 **取值范围：** ● 最小值：**-2** ● 最大值：**2** **默认取值：** **0（表示该参数未生效）**
frequency_penalty	否	Float	**参数解释：** 用于调整模型对频繁出现的Token的处理方式。即如果一个Token在训练集中出现的频率较高，那么模型在生成这个Token时会受到一定的惩罚。当frequency_penalty的值为正数时，模型会更倾向于生成出现频率较低的Token，即模型会更倾向于使用不常见的词汇。 **约束限制：** 不涉及 **取值范围：** 最小值：**-2** 最大值：**2** **默认取值：** **0（表示该参数未生效）**

表 3-7 ChatCompletionMessageParam

参数	是否必选	参数类型	描述
role	是	String	**参数解释：** 对话的角色，默认取值范围：system、user、assistant、tool、function。支持自定义。 如果需要模型以某个人设形象回答问题，可以将role参数设置为system。不使用人设时，可设置为user。 返回参数时，为固定值：assistant。 在一次会话请求中，人设只需要设置一次。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** system、user、assistant、tool、function
content	是	String	**参数解释：** 对话的内容，可以是任意文本，单位token。 **约束限制：** 设置多轮对话时，message中content个数不能超过20。 最小长度：**1** 最大长度：不同模型支持的token长度。 **取值范围：** 不涉及 **默认取值：** **None**

响应参数

非流式

状态码： 200

表 3-8 响应 Body 参数

参数	参数类型	描述
id	String	**参数解释：** 用来标识每个响应的唯一字符串。 **约束限制：** 形式为："chatcmpl-{random_uuid()}"。 **取值范围：** 不涉及 **默认取值：** 不涉及
object	String	**参数解释：** 固定为"chat.completion"。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
created	Integer	**参数解释：** 响应生成的时间，单位：s。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
model	String	**参数解释：** 请求模型ID。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

参数	参数类型	描述
choices	Array of **ChatCompletionResponseChoice** objects	**参数解释：** 生成的文本列表。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
usage	**UsageInfo** object	**参数解释：** 该对话请求的token用量信息。该参数可以帮助用户了解和控制模型的使用情况，避免超出Tokens限制。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
prompt_logprobs	Object	**参数解释：** 输入文本以及对应token的对数概率信息。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** **null**

表 3-9 ChatCompletionResponseChoice

参数	参数类型	描述
message	ChatMessage object	**参数解释：** 生成的文本内容。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
index	Integer	**参数解释：** 生成的文本在列表中的索引，从0开始。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
finish_reason	String	**参数解释：** 模型停止生成token的原因。 **约束限制：** 不涉及 **取值范围：** [stop, length, content_filter, tool_calls, insufficient_system_resource] ● stop：模型自然停止生成，或遇到stop序列中列出的字符串。 ● length：输出长度达到了模型上下文长度限制，或达到了max_tokens的限制。 ● content_filter：输出内容因触发过滤策略而被过滤。 ● tool_calls：模型决定调用外部工具（函数/API）来完成任务。 ● insufficient_system_resource：系统推理资源不足，生成被打断。 **默认取值：** **stop**

参数	参数类型	描述
logprobs	Object	**参数解释：** 评估指标，表示推理输出的置信度。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** **null**
stop_reason	Union[Integer, String]	**参数解释：** 导致生成停止的token id或者字符串。如果是遇到EOS token则返回默认值。如果是因为用户请求参数中指定的stop参数中的字符串或者token id，则返回对应的字符串或者token id。不是openAI接口标准字段，但vllm接口支持。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** **None**

表 3-10 UsageInfo

参数	参数类型	描述
prompt_tokens	Number	**参数解释：** 用户prompt中所包含的Token数。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

参数	参数类型	描述
total_tokens	Number	**参数解释：** 该次对话请求中，所有Token的数量。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
completion_tokens	Number	**参数解释：** 推理模型所产生的答案的Token数量。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

表 3-11 ChatMessage

参数	参数类型	描述
role	String	**参数解释：** 生成这条消息的角色。固定为：assistant。 **约束限制：** 不涉及 **取值范围：** assistant **默认取值：** assistant
content	String	**参数解释：** 对话的内容。 **约束限制：** 最小长度：**1** 最大长度：不同模型支持的token长度。 **取值范围：** 不涉及 **默认取值：** 不涉及

参数	参数类型	描述
reasoning_content	String	**参数解释：** 内容为在最终答案之前的推理内容（模型的思考过程）。 **约束限制：** 仅适用于DeepSeek-R1模型。 **取值范围：** 不涉及 **默认取值：** 不涉及

流式（stream参数为true）

状态码： 200

表 3-12 流式输出的数据单元

参数	参数类型	描述
data	CompletionStreamResponse object	**参数解释：** stream=true时，模型生成的消息以流式形式返回。生成的内容以增量的方式逐步发送回来，每个data字段均包含一部分生成的内容，直到所有data返回，响应结束。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

表 3-13 CompletionStreamResponse

参数	参数类型	描述
id	String	**参数解释：** 该对话的唯一标识符。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
created	Integer	**参数解释：** 创建聊天完成时的Unix时间戳（以秒为单位）。流式响应的每个chunk的时间戳相同。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
model	String	**参数解释：** 生成该completion的模型名。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
object	String	**参数解释：** 对象的类型，其值为chat.completion.chunk。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

参数	参数类型	描述
choices	ChatCompletionResponseStreamChoice	**参数解释：** 模型生成的completion的选择列表。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

表 3-14 ChatCompletionResponseStreamChoice

参数	参数类型	描述
index	Integer	**参数解释：** 该completion在模型生成的completion的选择列表中的索引。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

参数	参数类型	描述
finish_reason	String	**参数解释：** 模型停止生成token的原因。 **约束限制：** 不涉及 **取值范围：** [stop, length, content_filter, tool_calls, insufficient_system_resource] ● stop：模型自然停止生成，或遇到stop序列中列出的字符串。 ● length：输出长度达到了模型上下文长度限制，或达到了max_tokens的限制。 ● content_filter：输出内容因触发过滤策略而被过滤。 ● tool_calls：模型决定调用外部工具（函数/API）来完成任务。 ● insufficient_system_resource：系统推理资源不足，生成被打断。 **默认取值：** 不涉及

状态码： 400

表 3-15 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码。
error_msg	String	错误信息。

请求示例

- 非流式

```
V1推理接口：
POST https://{endpoint}/v1/{project_id}/alg-infer/3rdnlp/service/{deployment_id}/v1/chat/completions

Request Header:
Content-Type: application/json
X-Auth-Token:
MIINRwYJKoZIhvcNAQcCoIINODCCDTQCAQExDTALBglghkgBZQMEAgEwgguVBgkqhkiG...

Request Body:
{
  "model":"DeepSeek-V3",
  "messages":[
    {
      "role":"user",
      "content":"你好"
```

```
    }]
}
V2推理接口：
POST https://{endpoint}/api/v2/chat/completions

Request Header:
Content-Type: application/json
Authorization: Bearer 201ca68f-45f9-4e19-8fa4-831e...

Request Body:
{
  "model":"DeepSeek-V3",
  "messages":[
    {
      "role":"user",
      "content":"你好"
    }]
}
```

- 流式（stream参数为true）

```
V1推理接口：
POST https://{endpoint}/v1/{project_id}/alg-infer/3rdnlp/service/{deployment_id}/v1/chat/completions

Request Header:
Content-Type: application/json
X-Auth-Token:
MIINRwYJKoZIhvcNAQcCoIINODCCDTQCAQExDTALBglghkgBZQMEAgEwgguVBgkqhkiG...

Request Body:
{
  "model":"DeepSeek-V3",
  "messages":[
    {
      "role":"user",
      "content":"你好"
    }],
  "stream":true
}
V2推理接口：
POST https://{endpoint}/api/v2/chat/completions

Request Header:
Content-Type: application/json
Authorization: Bearer 201ca68f-45f9-4e19-8fa4-831e...

Request Body:
{
  "model":"DeepSeek-V3",
  "messages":[
    {
      "role":"user",
      "content":"你好"
    }],
  "stream":true
}
```

响应示例

状态码： 200

OK

- 非流式问答响应

```
{
    "id": "chat-9a75fc02e45d48db94f94ce38277beef",
    "object": "chat.completion",
    "created": 1743403365,
    "model": "DeepSeek-V3",
    "choices": [
```



```
        {
            "index": 0,
            "message": {
                "role": "assistant",
                "content": "你好！有什么我可以帮助你的吗？",
                "tool_calls": []
            },
            "finish_reason": "stop"
        }
    ],
    "usage": {
        "prompt_tokens": 64,
        "total_tokens": 73,
        "completion_tokens": 9
    }
}
```


- 带有思维链的非流式问答响应


```
{
    "id": "81c34733-0e7c-4b4b-a044-1e1fcd54b8db",
    "model": "deepseek-r1_32k",
    "created": 1747485310,
    "choices": [
        {
            "index": 0,
            "message": {
                "role": "assistant",
                "content": "\n\n你好！很高兴见到你，有什么我可以帮忙的吗？",
                "reasoning_content": "嗯，用户刚刚发了一个简短的“你好”，这是在用中文打招呼。首先我需要确认他们的需求是什么，可能只是想测试一下回复，或者有具体的问题要问。另外，我需要考虑是否需要用英文回应，但用户用了中文，用中文回复更合适吧。\n\n然后，我要确保回复友好且符合指南，不能涉及敏感内容。用户可能期待进一步的对话或者有问题需要帮助。这时候应该保持开放式的回答，邀请他们提出具体的问题或需求。比如，可以说“你好！很高兴见到你，有什么我可以帮忙的吗？”这样既礼貌又主动提供帮助。\n\n另外，注意避免使用任何格式或markdown，保持自然简洁。可能存在用户刚接触这个平台，不熟悉如何提问的情况，所以用鼓励的语气可能会更好。检查有没有任何拼写或语法错误，确保回复正确无误。\n",
                "tool_calls": [
                ]
            },
            "finish_reason": "stop"
        }
    ],
    "usage": {
        "completion_tokens": 184,
        "prompt_tokens": 6,
        "total_tokens": 190
    }
}
```


- 流式问答响应


```
V1推理接口返回体：
data:
{"id":"chat-97313a4bc0a342558364345de0380291","object":"chat.completion.chunk","created":1743404317,"model":"DeepSeek-V3","choices":[{"index":0,"message":
{"role":"assistant"},"logprobs":null,"finish_reason":null}]

data:
{"id":"chat-97313a4bc0a342558364345de0380291","object":"chat.completion.chunk","created":1743404317,"model":"DeepSeek-V3","choices":[{"index":0,"message":{"content":"你好
"},"logprobs":null,"finish_reason":null}]}

data:
{"id":"chat-97313a4bc0a342558364345de0380291","object":"chat.completion.chunk","created":1743404317,"model":"DeepSeek-V3","choices":[{"index":0,"message":{"content":"，有什么我能帮您的吗?
"},"logprobs":null,"finish_reason":"stop","stop_reason":null}]}

data:[DONE]
V2推理接口返回体：
data:
{"id":"chat-97313a4bc0a342558364345de0380291","object":"chat.completion.chunk","created":174340
```


4317,"model":"DeepSeek-V3","choices":[{"index":0,"delta":
{"role":"assistant"},"logprobs":null,"finish_reason":null}]

data:
{"id":"chat-97313a4bc0a342558364345de0380291","object":"chat.completion.chunk","created":174340
4317,"model":"DeepSeek-V3","choices":[{"index":0,"delta":{"content":"你好
"},"logprobs":null,"finish_reason":null}]}

data:
{"id":"chat-97313a4bc0a342558364345de0380291","object":"chat.completion.chunk","created":174340
4317,"model":"DeepSeek-V3","choices":[{"index":0,"delta":{"content":"，有什么我能帮您的吗?
"},"logprobs":null,"finish_reason":"stop","stop_reason":null}]}

data:[DONE]

- 带有思维链的流式问答响应

V1推理接口返回体:
data:{"id":"chat-
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"
:"DeepSeek-R1","choices":[{"index":0,"message":
{"role":"assistant","content":""},"logprobs":null,"finish_reason":null}],"usage":
{"prompt_tokens":6,"total_tokens":6,"completion_tokens":0}}

data:{"id":"chat-
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"
:"DeepSeek-R1","choices":[{"index":0,"message":{"reasoning_content":"嗯
"},"logprobs":null,"finish_reason":null}],"usage":
{"prompt_tokens":6,"total_tokens":7,"completion_tokens":1}}

data:{"id":"chat-
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"
:"DeepSeek-R1","choices":[{"index":0,"message":{"reasoning_content":",
"},"logprobs":null,"finish_reason":null}],"usage":
{"prompt_tokens":6,"total_tokens":8,"completion_tokens":2}}

data:{"id":"chat-
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"
:"DeepSeek-R1","choices":[{"index":0,"message":{"reasoning_content":"用户发
"},"logprobs":null,"finish_reason":null}],"usage":
{"prompt_tokens":6,"total_tokens":10,"completion_tokens":4}}

...

data:{"id":"chat-
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"
:"DeepSeek-R1","choices":[{"index":0,"message":{"reasoning_content":"生成
"},"logprobs":null,"finish_reason":null}],"usage":
{"prompt_tokens":6,"total_tokens":185,"completion_tokens":179}}

data:{"id":"chat-
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"
:"DeepSeek-R1","choices":[{"index":0,"message":{"reasoning_content":"最终的
"},"logprobs":null,"finish_reason":null}],"usage":
{"prompt_tokens":6,"total_tokens":186,"completion_tokens":180}}

data:{"id":"chat-
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"
:"DeepSeek-R1","choices":[{"index":0,"message":{"reasoning_content":"回复。
\n"},"logprobs":null,"finish_reason":null}],"usage":
{"prompt_tokens":6,"total_tokens":188,"completion_tokens":182}}

data:{"id":"chat-
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"
:"DeepSeek-R1","choices":[{"index":0,"message":{"content":"\n\n你好
"},"logprobs":null,"finish_reason":null}],"usage":
{"prompt_tokens":6,"total_tokens":191,"completion_tokens":185}}

data:{"id":"chat-
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"

:"DeepSeek-R1","choices":[{"index":0,"message":{"content":"！很高兴
"},"logprobs":null,"finish_reason":null}],"usage":
{"prompt_tokens":6,"total_tokens":193,"completion_tokens":187}}

data:{"id":"chat-
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"
:"DeepSeek-R1","choices":[{"index":0,"message":{"content":"见到
"},"logprobs":null,"finish_reason":null}],"usage":
{"prompt_tokens":6,"total_tokens":194,"completion_tokens":188}}

data:{"id":"chat-
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"
:"DeepSeek-R1","choices":[{"index":0,"message":{"content":"你
"},"logprobs":null,"finish_reason":null}],"usage":
{"prompt_tokens":6,"total_tokens":195,"completion_tokens":189}}

data:{"id":"chat-
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"
:"DeepSeek-R1","choices":[{"index":0,"message":{"content":"，有什么
"},"logprobs":null,"finish_reason":null}],"usage":
{"prompt_tokens":6,"total_tokens":197,"completion_tokens":191}}

data:{"id":"chat-
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"
:"DeepSeek-R1","choices":[{"index":0,"message":{"content":"我可以帮
"},"logprobs":null,"finish_reason":null}],"usage":
{"prompt_tokens":6,"total_tokens":199,"completion_tokens":193}}

data:{"id":"chat-
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"
:"DeepSeek-R1","choices":[{"index":0,"message":{"content":"您的吗
"},"logprobs":null,"finish_reason":null}],"usage":
{"prompt_tokens":6,"total_tokens":201,"completion_tokens":195}}

data:{"id":"chat-
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"
:"DeepSeek-R1","choices":[{"index":0,"message":{"content":"?
"},"logprobs":null,"finish_reason":"stop","stop_reason":null}],"usage":
{"prompt_tokens":6,"total_tokens":203,"completion_tokens":197}}

data:{"id":"chat-
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"
:"DeepSeek-R1","choices":[],"usage":{"prompt_tokens":6,"total_tokens":203,"completion_tokens":197}}

data:[DONE]
V2推理接口返回体：
data:{"id":"chat-
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"
:"DeepSeek-R1","choices":[{"index":0,"delta":
{"role":"assistant","content":""},"logprobs":null,"finish_reason":null}],"usage":
{"prompt_tokens":6,"total_tokens":6,"completion_tokens":0}}

data:{"id":"chat-
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"
:"DeepSeek-R1","choices":[{"index":0,"delta":{"reasoning_content":"嗯
"},"logprobs":null,"finish_reason":null}],"usage":
{"prompt_tokens":6,"total_tokens":7,"completion_tokens":1}}

data:{"id":"chat-
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"
:"DeepSeek-R1","choices":[{"index":0,"delta":{"reasoning_content":"，
"},"logprobs":null,"finish_reason":null}],"usage":
{"prompt_tokens":6,"total_tokens":8,"completion_tokens":2}}

data:{"id":"chat-
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"
:"DeepSeek-R1","choices":[{"index":0,"delta":{"reasoning_content":"用户发
"},"logprobs":null,"finish_reason":null}],"usage":
{"prompt_tokens":6,"total_tokens":10,"completion_tokens":4}}


```
...

data:{"id":"chat-
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"
:"DeepSeek-R1","choices":[{"index":0,"delta":{"reasoning_content":"生成
"},"logprobs":null,"finish_reason":null}],"usage":
{"prompt_tokens":6,"total_tokens":185,"completion_tokens":179}}

data:{"id":"chat-
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"
:"DeepSeek-R1","choices":[{"index":0,"delta":{"reasoning_content":"最终的
"},"logprobs":null,"finish_reason":null}],"usage":
{"prompt_tokens":6,"total_tokens":186,"completion_tokens":180}}

data:{"id":"chat-
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"
:"DeepSeek-R1","choices":[{"index":0,"delta":{"reasoning_content":"回复。
\n"},"logprobs":null,"finish_reason":null}],"usage":
{"prompt_tokens":6,"total_tokens":188,"completion_tokens":182}}

data:{"id":"chat-
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"
:"DeepSeek-R1","choices":[{"index":0,"delta":{"content":"\n\n你好
"},"logprobs":null,"finish_reason":null}],"usage":
{"prompt_tokens":6,"total_tokens":191,"completion_tokens":185}}

data:{"id":"chat-
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"
:"DeepSeek-R1","choices":[{"index":0,"delta":{"content":"！很高兴
"},"logprobs":null,"finish_reason":null}],"usage":
{"prompt_tokens":6,"total_tokens":193,"completion_tokens":187}}

data:{"id":"chat-
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"
:"DeepSeek-R1","choices":[{"index":0,"delta":{"content":"见到
"},"logprobs":null,"finish_reason":null}],"usage":
{"prompt_tokens":6,"total_tokens":194,"completion_tokens":188}}

data:{"id":"chat-
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"
:"DeepSeek-R1","choices":[{"index":0,"delta":{"content":"你
"},"logprobs":null,"finish_reason":null}],"usage":
{"prompt_tokens":6,"total_tokens":195,"completion_tokens":189}}

data:{"id":"chat-
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"
:"DeepSeek-R1","choices":[{"index":0,"delta":{"content":"，有什么
"},"logprobs":null,"finish_reason":null}],"usage":
{"prompt_tokens":6,"total_tokens":197,"completion_tokens":191}}

data:{"id":"chat-
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"
:"DeepSeek-R1","choices":[{"index":0,"delta":{"content":"我可以帮
"},"logprobs":null,"finish_reason":null}],"usage":
{"prompt_tokens":6,"total_tokens":199,"completion_tokens":193}}

data:{"id":"chat-
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"
:"DeepSeek-R1","choices":[{"index":0,"delta":{"content":"您的吗
"},"logprobs":null,"finish_reason":null}],"usage":
{"prompt_tokens":6,"total_tokens":201,"completion_tokens":195}}

data:{"id":"chat-
cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model"
:"DeepSeek-R1","choices":[{"index":0,"delta":{"content":"?
"},"logprobs":null,"finish_reason":"stop","stop_reason":null}],"usage":
{"prompt_tokens":6,"total_tokens":203,"completion_tokens":197}}
```

```
data:{"id":"chat-cc897cfa872a4fc993a803bbddf9268a","object":"chat.completion.chunk","created":1747485542,"model":"DeepSeek-R1","choices":[],"usage":{"prompt_tokens":6,"total_tokens":203,"completion_tokens":197}}

data:[DONE]
```

- 流式问答，内容审核不通过时的响应

```
event:moderation data:{"suggestion":"block","reply":"作为AI语言模型，我的目标是以积极、正向和安全的方式提供帮助和信息，您的问题超出了我的回答范围。"}

data:[DONE]
```

状态码

请参见**状态码**。

错误码

请参见**错误码**。

3.1.1.2 Qwen 三方 VL 大模型

功能介绍

Qwen2.5-VL系列模型，具备图像识别、精准视觉定位、文字识别和理解、文档解析、视频理解等能力。

URI

多模态推理服务提供两种推理接口调用：

- 盘古推理接口（V1推理接口）
- 业界通用的OpenAI格式接口（V2推理接口）

V1、V2调用接口的鉴权方式不同，请求体和返回体略有差异。两种接口定义如**推理接口**所示。

表 3-16 推理接口

API分类	API访问路径（URI）
V1推理接口	POST /v1/{project_id}/deployments/{deployment_id}/chat/completions
V2推理接口	POST /api/v2/chat/completions

V1推理接口URI需要输入额外参数，参数说明如**路径参数**所示：

表 3-17 路径参数

参数	是否必选	参数类型	描述
project_id	是	String	**参数解释：** 项目ID，获取方法请参见**获取项目ID**。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
deployment_id	是	String	**参数解释：** 模型的部署ID，获取方法请参见**获取模型部署ID**。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

请求参数

V1、V2推理接口的鉴权方式不同，请求参数与响应参数也有不同，说明如下：

Header参数

1. V1接口支持Token鉴权方式，也支持API Key鉴权方式。两种鉴权方式请求Header参数说明如下：
   - 使用**Token认证**方式的请求Header参数见**请求Header参数（Token认证）**。

表 3-18 请求 Header 参数（Token 认证）

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	**参数解释：** 用户Token。 用于获取操作API的权限。如**Token认证**中响应消息头中X-Subject-Token的值即为Token。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
Content-Type	是	String	**参数解释：** 发送的实体的MIME类型。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** application/json

- 使用**API Key认证**方式的请求Header参数见**请求Header参数（API Key认证）**。

表 3-19 请求 Header 参数（API Key 认证）

参数	是否必选	参数类型	描述
X-Apig-AppCode	是	String	**参数解释：** API Key值。 用于获取操作API的权限。API Key认证响应消息头中X-Apig-AppCode的值即为API Key。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
Content-Type	是	String	**参数解释：** 发送的实体的MIME类型。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** application/json

2. V2接口仅支持API Key鉴权方式。请求Header参数见表3-20

表 3-20 请求 Header 参数（OpenAI 格式的 API Key 认证）

参数	是否必选	参数类型	描述
Authorization	是	String	**参数解释：** 用户创建应用接入获取的API Key，拼接“Bearer ”后的字符串。示例：Bearer d59******9C3。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

参数	是否必选	参数类型	描述
Content-Type	是	String	**参数解释：** 发送的实体的MIME类型。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** application/json

请求Body参数

V1、V2推理接口请求Body参数一致，如表3-21描述。

表 3-21 请求 Body 参数

参数	是否必选	参数类型	描述
messages	是	Array of **message** objects	**参数解释：** 多轮对话问答对。 **约束限制：** 不涉及 **取值范围：** 数组长度：**1 - 20** **默认取值：** 不涉及
model	V1推理接口：否 V2推理接口：是	String	**参数解释：** 使用的推理服务模型名称，为推理服务部署时指定的Deployed_Model，可在推理服务详情页面查询到。V2推理接口必须指定此参数，V1推理接口不需要此参数。 **约束限制：** 不涉及 **取值范围：** 字符串长度最大64，最小1。 **默认取值：** 不涉及

参数	是否必选	参数类型	描述
stream	否	Boolean	**参数解释：** 流式调用的开启开关。 **约束限制：** 不涉及 **取值范围：** ● **true**：开启流式调用 ● **false**：关闭流式调用 **默认取值：** **false**
temperature	否	Float	**参数解释：** 用于控制生成文本的多样性和创造力。参数的取值范围是0到1，其中0表示最低的随机性。一般来说，temperature越低，适合完成确定性的任务。temperature越高，如0.9，适合完成创造性的任务。temperature参数可以影响语言模型输出的质量和多样性，但也不是唯一的因素。还有其他一些参数，如top_p参数也可以用来调整语言模型的行为和偏好，但不建议同时更改这两个参数。 **约束限制：** 不涉及 **取值范围：** 最小值：**0** 最大值：**1** **默认取值：** 缺省值：**0.3**

参数	是否必选	参数类型	描述
top_p	否	Float	**参数解释：** 一种替代温度采样的方法，称为nucleus sampling，其中模型考虑具有top_p概率质量的标记的结果。通常建议更改此值或温度，但不要同时更改两者。通常建议更改top_p或temperature来调整生成文本的倾向性，但不要同时更改这两个参数。 **约束限制：** 不涉及 **取值范围：** [0, 1] **默认取值：** 缺省值：**0**
max_tokens	否	Integer	**参数解释：** 用于控制聊天回复的长度和质量。一般来说，较大的max_tokens值可以生成较长和较完整的回复，但也可能增加生成无关或重复内容的风险。较小的max_tokens值可以生成较短和较简洁的回复，但也可能导致生成不完整或不连贯的内容。因此，需要根据不同的场景和需求来选择合适的max_tokens值。 **约束限制：** 最小值为**1** **取值范围：** 不涉及 **默认取值：** 不涉及

参数	是否必选	参数类型	描述
presence_penalty	否	Float	**参数解释：** 用于控制生成文本中的重复程度。正值会根据它们到目前为止在文本中的现有频率来惩罚新tokens，从而降低模型逐字重复同一行的可能性。presence_penalty参数可以用来提高生成文本的多样性和创造性，避免生成单调或重复的内容。 **约束限制：** 不涉及 **取值范围：** 最小值：**-2** 最大值：**2** **默认取值：** 缺省值：**0**
frequency_penalty	否	Float	**参数解释：** 重复采样惩罚值，避免文本重复生成。 **约束限制：** 不涉及 **取值范围：** 最小值：**-2** 最大值：**2** **默认取值：** **0**

表 3-22 message

参数	是否必选	参数类型	描述
role	V1推理接口：否 V2推理接口：是	String	**参数解释：** 对话的角色，取值为system、user、assistant。 如果需要模型以某个人设形象回答问题，可以将role参数设置为system。不使用人设时，可设置为user。在一次会话请求中，人设只需要设置一次。多轮对话中，用户输入提示词的role设置为user，推理结果的role设置为assistant。 **约束限制：** 不涉及 **取值范围：** [system, user, assistant] **默认取值：** 不涉及
content	是	Array of **content** objects	**参数解释：** 问答对文本内容。 **约束限制：** 最小长度：**1** **取值范围：** 不涉及 **默认取值：** 不涉及

表 3-23 content

参数	是否必选	参数类型	描述
type	是	String	**参数解释：** 输入内容的类型。 **约束限制：** 不涉及 **取值范围：** • text：文本 • image_url：图像 **默认取值：** 不涉及

参数	是否必选	参数类型	描述
text	否	String	**参数解释：** 问答对文本内容。 **约束限制：** 最小长度：**1** type为text时必传。 **取值范围：** 不涉及 **默认取值：** 不涉及
image_url	否 text、image_url不能同时为空	**image_url** object	**参数解释：** 问答对图像内容。 **约束限制：** type为image_url时必传。 **取值范围：** 不涉及 **默认取值：** 不涉及

表 3-24 image_url

参数	是否必选	参数类型	描述
url	是	String	**参数解释**: 标识符 + 图片的base64编码组成的字符串。 **约束限制：** 需要符合"{base64_str}"的格式，base64_str是图片的base64编码，示例：.....qkf/z。 **取值范围：** 不涉及 **默认取值：** 不涉及

响应参数

非流式响应（请求中stream参数为空或false）

状态码： 200

表 3-25 响应 Body 参数

参数	参数类型	描述
id	String	**参数解释：** 响应ID。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
created	Integer	**参数解释：** 响应时间。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
choices	Array of **ChatChoice** objects	**参数解释：** 模型回复。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
usage	**CompletionUsage** object	**参数解释：** tokens数量统计对象。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

表 3-26 ChatChoice

参数	参数类型	描述
index	Integer	**参数解释：** 回复的索引。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
message	Array of **MessageItem** objects	**参数解释：** 模型响应。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

表 3-27 MessageItem

参数	参数类型	描述
role	String	**参数解释：** 角色。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
content	String	**参数解释：** 模型响应。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

表 3-28 CompletionUsage

参数	参数类型	描述
completion_tokens	Number	**参数解释：** 表示模型生成的答案中包含的Token的数量。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
prompt_tokens	Number	**参数解释：** 表示生成结果时使用的提示文本的Tokens的数量。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
total_tokens	Number	**参数解释：** 对话过程中使用的Tokens总数。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

流式响应（请求中stream参数为true）

状态码： 200

表 3-29 流式输出的数据单元

参数	参数类型	描述
data	CompletionStreamResponse	**参数解释：** stream=true时，模型生成的消息以流式形式返回。生成的内容以增量的方式逐步发送回来，每个data字段均包含一部分生成的内容，直到所有data返回，响应结束。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

表 3-30 CompletionStreamResponse

参数	参数类型	描述
id	String	**参数解释：** 该对话的唯一标识符。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
created	Integer	**参数解释：** 创建聊天完成时的Unix时间戳（以秒为单位）。流式响应的每个chunk的时间戳相同。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

参数	参数类型	描述
model	String	**参数解释：** 生成该completion的模型名。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
object	String	**参数解释：** 对象的类型, 其值为chat.completion.chunk。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
choices	ChatCompletionResponseStreamChoice	**参数解释：** 模型生成的completion的选择列表。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
usage	UsageInfo	**参数解释：** 该对话请求的token用量信息。该参数可以帮助用户了解和控制模型的使用情况，避免超出Tokens限制。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

表 3-31 ChatCompletionResponseStreamChoice

参数	参数类型	描述
index	Integer	**参数解释：** 该completion在模型生成的completion的选择列表中的索引。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
finish_reason	String	**参数解释：** 模型停止生成token的原因。 **约束限制：** 不涉及 **取值范围：** [stop, length, content_filter, tool_calls, insufficient_system_resource] ● stop：模型自然停止生成，或遇到stop序列中列出的字符串。 ● length ：输出长度达到了模型上下文长度限制，或达到了max_tokens的限制。 ● content_filter：输出内容因触发过滤策略而被过滤。 ● tool_calls：模型决定调用外部工具（函数/API）来完成任务。 ● insufficient_system_resource：系统推理资源不足，生成被打断。 **默认取值：** 不涉及
delta	DeltaMessage	**参数解释：** V2推理接口流式返回的一个completion增量。 V1推理接口返回体不包含此参数。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

参数	参数类型	描述
message	DeltaMessage	**参数解释：** V1推理接口流式返回的一个completion增量。 V2推理接口返回体不包含此参数。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

表 3-32 DeltaMessage

参数	参数类型	描述
role	String	**参数解释：** 产生这条消息的角色。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
content	String	**参数解释：** completion增量的内容。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

参数	参数类型	描述
reasoning_content	String	**参数解释：** 内容为最终答案之前的推理内容（模型的思考过程）。 **约束限制：** 仅适用于支持思考过程的模型。 **取值范围：** 不涉及 **默认取值：** 不涉及

表 3-33 UsageInfo

参数	参数类型	描述
prompt_tokens	Integer	**参数解释：** 用户输入的提示词及默认人设的Token数。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
completion_tokens	Integer	**参数解释：** 推理服务返回结果的Token数。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
total_tokens	Integer	**参数解释：** 总消耗Token数。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

状态码： 400

接口报错的场景下，V1推理接口返回的报错信息符合华为云规范；V2推理接口则会对外透传推理服务返回的错误信息，通常符合OpenAi接口格式。

表 3-34 响应 Body 参数

参数	参数类型	描述
error_msg	String	错误信息。
error_code	String	错误码。
details	List<Object>	推理服务返回的报错信息，具体的格式、内容取决于推理服务。

表 3-35 V2 推理接口响应错误信息 Body 参数

参数	参数类型	描述
error	**ErrorResp**	错误信息。
id	String	请求ID。

表 3-36 ErrorResp

参数	参数类型	描述
code	String	错误码。
type	String	错误类型。
message	String	错误详情。

请求示例

接口URL与消息头：

```
V1推理接口：
POST https://mastudio.cn-southwest-2.myhuaweicloud.com/v1/{project_id}/deployments/{deployment_id}/chat/completions

Request Header:
Content-Type: application/json
X-Auth-Token: MIINRwYJKoZIhvcNAQcCoIINODCCDTQCAQExDTALBglghkgBZQMEAgEwgguVBgkqhkiG...

V2推理接口：
POST https://mastudio.cn-southwest-2.myhuaweicloud.com/api/v2/chat/completions

Request Header:
Content-Type: application/json
Authorization: Bearer 201ca68f-45f9-4e19-8fa4-831e...
```

请求体示例：

```
{
   "temperature": 0.5,
   "model": "Qwen25-vl-32b", // 仅V2接口需要此参数
   "messages": [
      {
         "role":"user", // 仅V2接口需要此参数
         "content": [
            {
               "type": "image_url",
               "image_url": {
                  "url": ".....qVKgqkf/Z"
               }
            },
            {
               "type": "text",
               "text": "图中有什么？"
            }
         ]
      }
   ],
   "presence_penalty": 0.5,
   "frequency_penalty": 0.5,
   "max_tokens": 2048,
   "stream": false
}
```

多轮问答请求示例：

```
{
   "temperature": 0.5,
   "model": "Qwen25-vl-32b",   // 仅V2接口需要此参数
   "messages": [{
         "role": "user",    // 仅V2接口需要此参数
         "content": [{
               "type": "image_url",
               "image_url": {
                  "url": ".....qVKgqkf/Z"
               }
            },
            {
               "type": "text",
               "text": "图中有什么？"
            }
         ]
      },
      {
         "role": "assistant",
         "content": [{
            "type": "text",
            "text": "这是一张飞机在天空中飞行的图片。它显示了飞机及其翼展和发动机的大小,以及它所穿越的空气量。这架飞机是一架军用喷气式战斗机,机身颜色为黑色,机头有一个大螺旋桨。背景中的云层表明飞机正在接近高空,很可能是在航程的中间。"
         }]
      },
      {
         "role": "user",   // 仅V2接口需要此参数
         "content": [{
               "type": "image_url",
               "image_url": {
                  "url": ".....qVKgqkf/Z"
               }
            },
            {
               "type": "text",
               "text": "这张图与第一张图有什么差异？"
            }
         ]
      }
   ],
   "presence_penalty": 0.5,
```



```
    "frequency_penalty": 0.5,
    "max_tokens": 2048,
    "stream": false
}
```

响应示例

状态码： 200

非流式问答响应示例：

```
{
    "id": "chat-38ea6118a5d14e38b7d592211bbd31a6",
    "object": "chat.completion",
    "created": 1749894390,
    "model": "Qwen25-vl-32b",
    "choices": [
        {
            "index": 0,
            "message": {
                "role": "assistant",
                "reasoning_content": null,
                "content": "这是一张飞机在天空中飞行的图片。它显示了飞机及其翼展和发动机的大小,以及它所穿越的空气量。这架飞机是一架军用喷气式战斗机,机身颜色为黑色,机头有一个大螺旋桨。背景中的云层表明飞机正在接近高空,很可能是在航程的中间。",
                "tool_calls": [
                ]
            },
            "logprobs": null,
            "finish_reason": "stop",
            "stop_reason": null
        }
    ],
    "usage": {
        "prompt_tokens": 3189,
        "total_tokens": 3236,
        "completion_tokens": 47
    },
    "prompt_logprobs": null
}
```

流式问答响应示例：

```
V1推理接口响应：
data:
{"id":"chat-59170add0fd1427bbca0388431058d45","object":"chat.completion.chunk","created":1745725837,"model":"Qwen25-vl-32b","choices":[{"index":0,"logprobs":null,"finish_reason":null,"message":{"role":"assistant"}}],"usage":{"prompt_tokens":64,"total_tokens":64,"completion_tokens":0}}

data:
{"id":"chat-59170add0fd1427bbca0388431058d45","object":"chat.completion.chunk","created":1745725837,"model":"Qwen25-vl-32b","choices":[{"index":0,"logprobs":null,"finish_reason":null,"message":{"content":"在这"}}],"usage":{"prompt_tokens":64,"total_tokens":65,"completion_tokens":1}}

data:
{"id":"chat-59170add0fd1427bbca0388431058d45","object":"chat.completion.chunk","created":1745725837,"model":"Qwen25-vl-32b","choices":
[{"index":0,"logprobs":null,"finish_reason":"stop","stop_reason":null,"message":{"content":"张"}}],"usage":
{"prompt_tokens":64,"total_tokens":73,"completion_tokens":2}}

data:
{"id":"chat-59170add0fd1427bbca0388431058d45","object":"chat.completion.chunk","created":1745725837,"model":"Qwen25-vl-32b","choices":
[{"index":0,"logprobs":null,"finish_reason":"stop","stop_reason":null,"message":{"content":"图片"}}],"usage":
{"prompt_tokens":64,"total_tokens":73,"completion_tokens":3}}

.....

data:
```


{"id":"chat-59170add0fd1427bbca0388431058d45","object":"chat.completion.chunk","created":1745725837,"
model":"pQwen25-vl-32b","choices":[],"usage":
{"prompt_tokens":64,"total_tokens":73,"completion_tokens":9}}

event:{"usage":{"completionTokens":9,"promptTokens":64,"totalTokens":73},"tokens":64,"token_number":9}

data:[DONE]

V2推理接口响应：

data:
{"id":"chat-59170add0fd1427bbca0388431058d45","object":"chat.completion.chunk","created":1745725837,"
model":"Qwen25-vl-32b","choices":[{"index":0,"logprobs":null,"finish_reason":null,"delta":
{"role":"assistant"}}],"usage":{"prompt_tokens":64,"total_tokens":64,"completion_tokens":0}}

data:
{"id":"chat-59170add0fd1427bbca0388431058d45","object":"chat.completion.chunk","created":1745725837,"
model":"Qwen25-vl-32b","choices":[{"index":0,"logprobs":null,"finish_reason":null,"delta":{"content":"在这
"}}],"usage":{"prompt_tokens":64,"total_tokens":65,"completion_tokens":1}}

data:
{"id":"chat-59170add0fd1427bbca0388431058d45","object":"chat.completion.chunk","created":1745725837,"
model":"Qwen25-vl-32b","choices":
[{"index":0,"logprobs":null,"finish_reason":"stop","stop_reason":null,"delta":{"content":"张"}}],"usage":
{"prompt_tokens":64,"total_tokens":73,"completion_tokens":2}}

data:
{"id":"chat-59170add0fd1427bbca0388431058d45","object":"chat.completion.chunk","created":1745725837,"
model":"Qwen25-vl-32b","choices":
[{"index":0,"logprobs":null,"finish_reason":"stop","stop_reason":null,"delta":{"content":"图片"}}],"usage":
{"prompt_tokens":64,"total_tokens":73,"completion_tokens":3}}

.....

data:
{"id":"chat-59170add0fd1427bbca0388431058d45","object":"chat.completion.chunk","created":1745725837,"
model":"Qwen25-vl-32b","choices":[],"usage":
{"prompt_tokens":64,"total_tokens":73,"completion_tokens":9}}

event:{"usage":{"completionTokens":9,"promptTokens":64,"totalTokens":73},"tokens":64,"token_number":9}

data:[DONE]

状态码

请参见**状态码**。

错误码

请参见**错误码**。

3.2 数据工程接口

3.2.1 查询数据血缘

功能介绍

客户通过obs导入原始数据集，可基于该obs路径查询所有基于该路径创建的原始数据集及后续的血缘信息。

URI

GET /v1/{project_id}/workspaces/{workspace_id}/data-management/lineages

请求参数

表 3-37 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	**参数解释：** 用户Token。 用于获取操作API的权限。如Token认证中响应消息头中X-Subject-Token的值即为Token。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
Content-Type	是	String	**参数解释：** 发送的实体的MIME类型。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** application/json

表 3-38 请求 Query 参数

参数	是否必选	参数类型	描述
limit	是	integer	**参数解释：** 接口返回的血缘数量上限。 **约束限制：** 不涉及。 **取值范围：** [1, 1000] **默认取值：** 100

参数	是否必选	参数类型	描述
from_path	是	string	**参数解释：** 来源obs路径。 **约束限制：** 最终租户桶下的OBS全路径。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

响应参数

参数	参数类型	描述
lineages	array	**参数解释：** 数据集血缘列表。 **约束限制：** 列表内的item为Lineage类型。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

请求示例

```
GET https://{endpoint}/v1/{project_id}/workspaces/{workspace_id}/data-management/lineages?
limit=100&from_path=bucket/folder1/folder2

Requet Header:
Content_Type: application/json
X-Auth-Token: MIIVV...

Request Params:
limit: 1000
from_path: bucket/folder1/folder2
```

响应示例

```
{
   "lineages": [
      {
         "id": null,
         "from_id": null,
         "from_name": null,
         "from_catalog": null,
         "from_type": "OBS",
         "to_id": "1352299121133883392",
         "to_name": null,
         "to_catalog": "ORIGINAL",
         "to_type": "DATASET",
```

```
            "process_id": null,
            "process_name": null,
            "process_type": null,
            "train_job_name": null,
            "model_type": null,
            "train_type": null,
            "create_time": null,
            "from_path": "bucket/folder",
            "from_path_existed": null
        },
        {
            "id": "1352299380551585793",
            "from_id": "1352299121133883392",
            "from_name": "时序-回归-test",
            "from_catalog": "ORIGINAL",
            "from_type": "DATASET",
            "to_id": "1352299379473649664",
            "to_name": "pub_时序回归",
            "to_catalog": "PUBLISH",
            "to_type": "DATASET",
            "process_id": "lt_97a2aa4cca744775aa5c7cfe3cb36121",
            "process_name": "pub_时序回归",
            "process_type": "PUBLISH",
            "train_job_name": null,
            "model_type": null,
            "train_type": null,
            "create_time": null,
            "from_path": null,
            "from_path_existed": null
        }
    ]
}
```


状态码

请参见**状态码**。

错误码

请参见**错误码**。

3.2.2 数据集彻底删除

功能介绍

只针对从obs上传的数据，在删除数据集的时候要关联删除OBS下对应的原始数据。

URI

POST /v1/{project_id}/workspaces/{workspace_id}/data-management/dataset/permanent-delete

请求参数

表 3-39 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	**参数解释：** 用户Token。 用于获取操作API的权限。如Token认证中响应消息头中X-Subject-Token的值即为Token。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
Content-Type	是	String	**参数解释：** 发送的实体的MIME类型。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** application/json

表 3-40 请求 Body 参数

参数	是否必选	参数类型	描述
dataset_name	是	string	**参数解释：** 数据集名称。 **约束限制：** 名称长度范围[1,128]。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

参数	是否必选	参数类型	描述
catalog	否	CatalogEnum	**参数解释：** 数据集形态。 **约束限制：** 不涉及。 **取值范围：** ● ORIGINAL：执行数据导入产生的数据集类型。 ● PROCESS：执行数据加工产生的数据集类型。 ● PUBLISH：执行数据发布产生的数据集类型。 **默认取值：** 不涉及。
delete_obs	否	boolean	**参数解释：** 删除obs数据。 **约束限制：** 不涉及。 **取值范围：** ● true：删除obs数据。 ● false：不删除obs数据。 **默认取值：** 不涉及。

响应参数

参数	参数类型	描述
dataset_name	string	**参数解释：** 数据集名称。 **约束限制：** 不涉及。 **取值范围：** 名称长度范围[1,128]。 **默认取值：** 不涉及。

参数	参数类型	描述
catalog	CatalogEnum	**参数解释：** 数据集形态。 **约束限制：** 不涉及。 **取值范围：** ● ORIGINAL：执行数据导入产生的数据集类型。 ● PROCESS：执行数据加工产生的数据集类型。 ● PUBLISH：执行数据发布产生的数据集类型。 **默认取值：** 不涉及。
result	boolean	**参数解释：** 操作结果。 **约束限制：** 不涉及。 **取值范围：** ● true：删除成功。 ● false：删除失败。 **默认取值：** 不涉及。

请求示例

彻底删除数据集对应的OBS原始数据

```
POST https://{endpoint}/v1/{project_id}/workspaces/{workspace_id}/data-management/dataset/permanent-delete

Requet Header:
Content_Type: application/json
X-Auth-Token: MIIVV...

Request Params:
dataset_name: pub_345135233
catalog: PROCESS
delete_obs:true
```

响应示例

```
{
   "DatasetOperationResp": [
      {
         "dataset_name": pub_345135233,
         "catalog": PROCESS,
         "result": true,
      },
```

```
    ]
}
```

状态码

请参见**状态码**。

错误码

请参见**错误码**。

3.3 Agent 应用接口

3.3.1 调用应用

功能介绍

通过调用创建好的应用API，输入问题，将得到应用执行的结果。

URI

POST /v1/{project_id}/agents/{agent_id}/conversations/{conversation_id}

获取URI方式请参见**请求URI**。

表 3-41 路径参数

参数	是否必选	参数类型	描述
project_id	是	String	**参数解释：** 项目ID，获取方法请参见**获取项目ID**。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

参数	是否必选	参数类型	描述
agent_id	是	String	**参数解释：** Agent ID，获取方式如下： 在“Agent开发”页面，左侧导航栏选择“工作台 > 应用”， 在所需Agent中单击“ ··· > 复制ID”。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
conversation_id	是	String	**参数解释：** 会话ID，唯一标识每个会话的标识符，可将会话ID设置为任意值，使用标准UUID格式。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

请求参数

表 3-42 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	**参数解释：** 用户Token。 用于获取操作API的权限。如**Token认证**中响应消息头中X-Subject-Token的值即为Token。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

参数	是否必选	参数类型	描述
Content-Type	是	String	**参数解释：** 发送的实体的MIME类型。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** application/json
stream	是	Boolean	**参数解释：** 是否开启流式调用，默认开启。 **约束限制：** 当前Agent只支持流式调用，需设置为true。 **取值范围：** ● true：开启。 ● false：不开启。 **默认取值：** 不涉及

表 3-43 请求 Body 参数

参数	是否必选	参数类型	描述
query	是	String	**参数解释：** 用户问题，作为运行Agent的输入。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

响应参数

流式（Header中的stream参数为true）

状态码： 200

表 3-44 流式输出的数据单元

参数	参数类型	描述
data	String	**参数解释：** ● stream=true时，执行Agent的消息以流式形式返回。 ● 生成的内容以增量的方式逐步发送回来，每个data字段均包含一部分生成的内容，直到所有data返回，响应结束。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

表 3-45 流式输出的数据单元

参数	参数类型	描述
event	String	**参数解释：** 数据单元类型。 **约束限制：** 不涉及 **取值范围：** ● start：开始节点，表示开始调用模型进行会话。 ● message：消息节点，表示模型返回的消息。 ● plugin_start：插件调用请求节点，表示调用插件的请求信息。 ● plugin_end：插件调用响应节点，表示调用插件的响应信息。 ● statistic_data：执行数据节点，包含本次调用的耗时信息。 ● summary_response：消息总结节点，包含本次调用的全量响应信息。 ● done：流式调用结束节点，表示流式响应结束。 **默认取值：** 不涉及

参数	参数类型	描述
content	Object	**参数解释：** 消息块内容，不同event的消息块内容不同。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
createdTime	long	**参数解释：** 消息块返回的时间戳，如1733817348963。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
latency	Object	**参数解释：** 耗时，包括以下三个元素： ● plugin：插件调用耗时。 ● model：模型调用耗时。 ● overall：总耗时。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
plugin	Object	**参数解释：** 插件请求信息，包括以下两个元素： ● name：插件名。 ● arguments：插件入参名。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

请求示例

流式（Header中的stream参数为true）

```
POST https://{endpoint}/v1/{project_id}/agent-run/agents/{agent_id}/conversations/{conversation_id}

Request Header:
Content-Type: application/json
X-Auth-Token: MIINRwYJKoZIhvcNAQcCoIINODCCDTQCAQExDTALBglghkgBZQMEAgEwgguVBgkqhkiG...
stream: true

Request Body:
{
  "query": "查询A12会议室在9:00到10:00的状态"
}
```

响应示例

```
data:{"event":"start","createdTime":1735558575017}

data:{"event":"message","content":"好的","createdTime":1735558576300}

data:{"event":"message","content":"，","createdTime":1735558576301}

data:{"event":"message","content":"我将","createdTime":1735558576301}

data:{"event":"message","content":"调用","createdTime":1735558576302}

data:{"event":"message","content":"query","createdTime":1735558576302}

data:{"event":"message","content":"_","createdTime":1735558576302}

data:{"event":"message","content":"meeting","createdTime":1735558576302}

data:{"event":"message","content":"_","createdTime":1735558576302}

data:{"event":"message","content":"room","createdTime":1735558576303}

data:{"event":"message","content":"_status","createdTime":1735558576303}

data:{"event":"message","content":"工具","createdTime":1735558576303}

data:{"event":"message","content":"来","createdTime":1735558576304}

data:{"event":"message","content":"查询","createdTime":1735558576304}

data:{"event":"message","content":"A","createdTime":1735558576304}

data:{"event":"message","content":"12","createdTime":1735558576304}

data:{"event":"message","content":"会议室","createdTime":1735558576305}

data:{"event":"message","content":"在","createdTime":1735558576305}

data:{"event":"message","content":"9","createdTime":1735558576305}

data:{"event":"message","content":":00","createdTime":1735558576305}

data:{"event":"message","content":"到","createdTime":1735558576306}

data:{"event":"message","content":"10","createdTime":1735558576306}

data:{"event":"message","content":":00","createdTime":1735558576306}

data:{"event":"message","content":"的状态","createdTime":1735558576306}
```

data:{"event":"message","content":"。","createdTime":1735558576306}

data:{"event":"message","content":"请","createdTime":1735558576307}

data:{"event":"message","content":"稍","createdTime":1735558576307}

data:{"event":"message","content":"等","createdTime":1735558576307}

data:{"event":"message","content":"。","createdTime":1735558576307}

data:{"event":"message","content":" ","createdTime":1735558576307}

data:{"event":"message","content":" query","createdTime":1735558576307}

data:{"event":"message","content":"_","createdTime":1735558576308}

data:{"event":"message","content":"meeting","createdTime":1735558576308}

data:{"event":"message","content":"_","createdTime":1735558576308}

data:{"event":"message","content":"room","createdTime":1735558576308}

data:{"event":"message","content":"_status","createdTime":1735558576308}

data:{"event":"message","content":"|","createdTime":1735558576308}

data:{"event":"message","content":"{\"","createdTime":1735558576309}

data:{"event":"message","content":"meeting","createdTime":1735558576309}

data:{"event":"message","content":"Room","createdTime":1735558576309}

data:{"event":"message","content":"\":","createdTime":1735558576309}

data:{"event":"message","content":"{\"","createdTime":1735558576309}

data:{"event":"message","content":"number","createdTime":1735558576310}

data:{"event":"message","content":"\":","createdTime":1735558576310}

data:{"event":"message","content":" 12","createdTime":1735558576310}

data:{"event":"message","content":"}","createdTime":1735558576310}

data:{"event":"message","content":",\"","createdTime":1735558576310}

data:{"event":"message","content":"start","createdTime":1735558576310}

data:{"event":"message","content":"\":\"","createdTime":1735558576311}

data:{"event":"message","content":"9","createdTime":1735558576311}

data:{"event":"message","content":":00","createdTime":1735558576311}

data:{"event":"message","content":"\",\"","createdTime":1735558576311}

data:{"event":"message","content":"end","createdTime":1735558576311}

data:{"event":"message","content":"\":\"","createdTime":1735558576311}

data:{"event":"message","content":"10","createdTime":1735558576311}

data:{"event":"message","content":":00","createdTime":1735558576312}

data:{"event":"message","content":"\"}","createdTime":1735558576312}

data:{"event":"message","content":" ","createdTime":1735558576312}

data:{"event":"plugin_start","type":"plugin","latency":{"overall":1.3},"plugin":

{"name":"query_meeting_room_status","arguments":"{\"meetingRoom\": {\"number\": 12}, \"start\": \"9:00\", \"end\": \"10:00\"}"},"createdTime":1735558576316}

data:{"event":"plugin_end","content":{"result":"空闲"},"role":"function","latency":{"overall":1.51,"plugin":0.0},"createdTime":1735558576521}

data:{"event":"start","createdTime":1735558576522}

data:{"event":"message","content":"A","createdTime":1735558576976}

data:{"event":"message","content":"12","createdTime":1735558576977}

data:{"event":"message","content":"会议室","createdTime":1735558576977}

data:{"event":"message","content":"在","createdTime":1735558576977}

data:{"event":"message","content":"9","createdTime":1735558576978}

data:{"event":"message","content":":00","createdTime":1735558576978}

data:{"event":"message","content":"到","createdTime":1735558576978}

data:{"event":"message","content":"10","createdTime":1735558576978}

data:{"event":"message","content":":00","createdTime":1735558576978}

data:{"event":"message","content":"的时间","createdTime":1735558576978}

data:{"event":"message","content":"段","createdTime":1735558576979}

data:{"event":"message","content":"内","createdTime":1735558576979}

data:{"event":"message","content":"是","createdTime":1735558576979}

data:{"event":"message","content":"空闲","createdTime":1735558576979}

data:{"event":"message","content":"的","createdTime":1735558576979}

data:{"event":"message","content":"。","createdTime":1735558576980}

data:{"event":"statistic_data","latency":{"overall":1.97},"createdTime":1735558576986}

data:{"event":"summary_response","content":"A12会议室在9:00到10:00的时间段内是空闲的。","role":"assistant","createdTime":1735558576987}

data:{"event":"done","createdTime":1735558577011}

状态码

请参见**状态码**。

错误码

请参见**错误码**。

3.3.2 调用工作流

功能介绍

通过调用创建好的工作流API，输入问题，将得到工作流执行的结果。

URI

POST /v1/{project_id}/workflows/{workflow_id}/conversations/{conversation_id}

获取URI方式请参见**请求URI**。

表 3-46 路径参数

参数	是否必选	参数类型	描述
project_id	是	String	**参数解释：** 项目ID，获取方法请参见**获取项目ID**。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
workflow_id	是	String	**参数解释：** Workflow ID，获取方式如下： 在“Agent开发”页面，左侧导航栏选择“工作台 > 工作流”，在所需工作流中单击“ *** > 复制ID”。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
conversation_id	是	String	**参数解释：** 会话ID，唯一标识每个会话的标识符，可将会话ID设置为任意值，使用标准UUID格式。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

请求参数

表 3-47 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	**参数解释：** 用户Token。 用于获取操作API的权限。如Token认证中响应消息头中X-Subject-Token的值即为Token。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
Content-Type	是	String	**参数解释：** 发送的实体的MIME类型。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** application/json
stream	否	Boolean	**参数解释：** 是否开启流式调用。 • true：开启。 • false：不开启。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

表 3-48 请求 Body 参数

参数	是否必选	参数类型	描述
inputs	是	Map<String, Object>	**参数解释：** 用户提出的问题，作为运行工作流的输入，与工作流开始节点输入参数对应。包含默认字段query为用户输入。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
plugin_configs	否	List<PluginConfig>	**参数解释：** 插件配置，当工作流有配置用户自定义插件节点时，可能需要配置鉴权信息等，具体结构定义详见表3-49。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

表 3-49 PluginConfig 参数

参数	是否必选	参数类型	描述
plugin_id	是	String	**参数解释：** 插件Id，获取方式如下： 在“Agent开发”页面，左侧导航栏选择“工作台 > 插件”， 在所需插件中单击“ *** > 复制ID”。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

参数	是否必选	参数类型	描述
config	是	Map<String, String>	**参数解释：** 插件配置信息。 ● 当工作流关联插件节点，并且插件是“用户级鉴权”时，需要在此配置对应的鉴权信息，例如针对如下插件，config可以配成：{"key2": "value"}。  ● 其他情况该参数无需传值，plugin_configs传空数组即可。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

响应参数

非流式（Header中的stream参数为false）

状态码： 200

表 3-50 非流式输出的数据单元

参数	参数类型	描述
outputs	Map<String, Object>	**参数解释：** 工作流最终输出内容，支持多个参数。 **说明** outputs示例如下： • responseContent参数是默认有的，值取为工作流结束节点里的“指定回复”内容。 • 支持用户在工作流结束节点的“输出参数”模块自定义配置参数，自定义配置参数将会显示在user_fields参数里。 "outputs":{"user_fields":{"aaa":"1","vvv":[{"role":"user","content":"1"}]},"responseContent":"你好！\uD83D\uDE0A 你输入了“1”，请问有什么我可以帮助你的吗？如果有具体问题或需求，随时告诉我哦！"}  **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
messages	List<Message>	**参数解释：** 工作流助手回复内容，如提问器节点问题消息，详见表3-51。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

参数	参数类型	描述
status	Map<String, Object>	**参数解释：** 状态，包含状态码code，状态描述desc。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
start_time	Long	**参数解释：** 开始时间。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
end_time	Long	**参数解释：** 结束时间。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

表 3-51 Message

参数	参数类型	描述
role	String	**参数解释：** 会话角色，支持user、assistant。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

参数	参数类型	描述
content	String	**参数解释：** 会话内容。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

流式（Header中的stream参数为true或不传）

状态码： 200

表 3-52 流式输出的数据单元

参数	参数类型	描述
data	String	**参数解释：** stream=true时，执行工作流的消息以流式形式返回。生成的内容以增量的方式逐步发送回来，每个data字段均包含一部分生成的内容，直到所有data返回，响应结束。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

表 3-53 流式输出的数据单元

参数	参数类型	描述
event	String	**参数解释：** 数据单元类型。 **约束限制：** 不涉及 **取值范围：** ● workflow_started：工作流开始事件，表示工作流开始运行。 ● workflow_finished：工作流结束事件，表示工作流结束运行。 ● message：消息事件，表示工作流执行过程中流式返回的消息。 ● error：错误事件，表示工作流执行错误信息。 ● end：结束事件，标识请求结束。 **默认取值：** 不涉及
data	Object	**参数解释：** 消息块内容，不同event的消息块内容不同。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

表 3-54 workflow_started 事件的数据单元

参数	参数类型	描述
start_time	Long	**参数解释：** 工作流开始时间。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

表 3-55 workflow_finished 事件的数据单元

参数	参数类型	描述
start_time	Long	**参数解释：** 工作流开始时间。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
end_time	Long	**参数解释：** 工作流结束时间。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
outputs	Map<String, Object>	**参数解释：** 工作流最终输出内容，支持多个参数。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
status	Map<String, Object>	**参数解释：** 状态，包含状态码code，状态描述desc。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

表 3-56 message 事件的数据单元

参数	参数类型	描述
text	String	**参数解释：** 工作流输出内容消息块。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
index	Integer	**参数解释：** 消息块索引。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
node_id	String	**参数解释：** 节点id。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
node_type	String	**参数解释：** 支持输出message事件的节点类型。 **约束限制：** 不涉及 **取值范围：** Message：消息节点。 End：结束节点。 Questioner：提问器节点。 Input：输入节点。 **默认取值：** 不涉及

参数	参数类型	描述
node_name	String	**参数解释：** 节点名。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

表 3-57 error 事件的数据单元

参数	参数类型	描述
code	String	**参数解释：** 工作流执行错误码。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
message	String	**参数解释：** 工作流执行错误消息。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
node_id	String	**参数解释：** 节点id。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

参数	参数类型	描述
node_type	String	**参数解释：** 发送错误事件的节点类型。 **约束限制：** 不涉及 **取值范围：** Start：开始节点。 End：结束节点。 LLM：大模型节点。 Workflow：工作流节点。 Agent：Agent节点。 Branch：判断节点。 IntentDetection：意图识别节点。 Code：代码节点。 Loop：循环节点。 Plugin：插件节点。 Mcp：MCP服务节点。 Message：消息节点。 Questioner：提问器节点。 Input：输入节点。 SetVariable：变量赋值节点。 Aggregation：变量聚合节点。 KnowledgeRepo：知识检索节点。 Unknown：未知节点。 **默认取值：** 不涉及
node_name	String	**参数解释：** 节点名。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

请求示例

```
POST https://{endpoint}/v1/{project_id}/agent-run/workflows/{workflow_id}/conversations/{conversation_id}

Request Header:
Content-Type: application/json
```

```
X-Auth-Token: MIINRwYJKoZIhvcNAQcCoIINODCCDTQCAQExDTALBglghkgBZQMEAgEwgguVBgkqhkiG...
stream: true
Request Body:
{
  "inputs": {
    "query": "你好"
  },
  "plugin_configs": [
    {
      "plugin_id": "xxxxxxxxx",
      "config": {
        "key": "value"
      }
    }
  ]
}
```

响应示例

非流式（Header中的stream参数为false）

输入节点返回：

```
{
    "conversation_id": "2c90493f-803d-431d-a197-57543d414317",
    "messages": [
        {
            "role": "assistant",
            "content": "{\"inputs\": [{\"actualType\": \"string\", \"sourceType\": \"null\", \"description\": \"姓名\",
\"name\": \"name\", \"type\": \"string\", \"required\": true}]}",
            "nodeId": "node_1745928389632",
            "nodeType": "Input",
            "nodeName": "输入"
        }

    ],
    "status": {
        "code": 3,
        "desc": "waiting"
    },
    "start_time": 1734336269313,
    "end_time": 1734336270908
}
```

提问器节点返回：

```
{
    "conversation_id": "f9a5540f-0c92-4f28-bd6e-f96ce04f5cc81",
    "messages": [
        {
            "role": "assistant",
            "content": "请您提供姓名, 年龄相关的信息",
            "nodeId": "node_1745929628452",
            "nodeType": "Questioner",
            "nodeName": "提问器"
        }
    ],
    "status": {
        "code": 3,
        "desc": "waiting"
    },
    "start_time": 1745929778250,
    "end_time": 1745929779951
}
```

结束节点返回：

```
{
    "conversation_id": "2c90493f-803d-431d-a197-57543d414317",
    "outputs": {
```



```
        "responseContent": "你好！有什么我可以帮助你的吗？"
    },
    "messages": [],
    "status": {
        "code": 1,
        "desc": "succeeded"
    },
    "start_time": 1734337068533,
    "end_time": 1734337082545
}
```


流式（Header中的stream参数为true或不传）

输入节点返回：


```
data:{"event":"workflow_started","data":{"start_time":1745929087614}}

data:{"event":"message","data":{"text":"{\"inputs\": [{\"actualType\": \"string\", \"sourceType\": \"null\", \"description\": \"姓名\", \"name\": \"name\", \"type\": \"string\", \"required\": true}]}","index":0,"node_id":"node_1745928389632","node_type":"Input","node_name":"输入"}}

data:{"event":"message","data":{"text":"","node_id":"node_1745928389632","node_type":"Input","node_name":"输入","is_finished":true}}

data:{"event":"end"}
```


提问器节点返回：


```
data:{"event":"workflow_started","data":{"start_time":1745929709955}}

data:{"event":"message","data":{"text":"请您提供姓名, 年龄相关的信息
","index":0,"node_id":"node_1745929628452","node_type":"Questioner","node_name":"提问器"}}

data:{"event":"message","data":{"text":"","node_id":"node_1745929628452","node_type":"Questioner","node_name":"提问器
","is_finished":true}}

data:{"event":"end"}
```


结束节点返回：


```
data:{"event":"workflow_started","data":{"start_time":1745929897770}}

data:{"event":"message","data":{"text":"","index":0,"node_id":"node_end","node_type":"End","node_name":"结束"}}

data:{"event":"message","data":{"text":"你好
","index":1,"node_id":"node_end","node_type":"End","node_name":"结束"}}

data:{"event":"message","data":{"text":"！
","index":2,"node_id":"node_end","node_type":"End","node_name":"结束"}}

data:{"event":"message","data":{"text":"有什么我可以帮助你的吗？
","index":3,"node_id":"node_end","node_type":"End","node_name":"结束"}}

data:{"event":"message","data":{"text":"","node_id":"node_end","node_type":"End","node_name":"结束
","is_finished":true}}

data:{"event":"workflow_finished","data":{"status":{"code":1,"desc":"succeeded"},"outputs":{"responseContent":"你好！有什么我可以帮助你的吗？
"},"start_time":1745929897770,"end_time":1745929898600}}

data:{"event":"end"}
```


状态码

请参见**状态码**。

错误码

请参见**错误码**。

3.4 Token 计算器

功能介绍

为了帮助用户更好地管理和优化Token消耗，平台提供了Token计算器工具。Token计算器可以帮助用户在模型推理前评估文本的Token数量，提供费用预估，并优化数据预处理策略。

URI

POST /v1/{project_id}/deployments/{deployment_id}/caltokens

获取URI方式请参见**请求URI**。

表 3-58 路径参数

参数	是否必选	参数类型	描述
project_id	是	String	**参数解释：** 项目ID，获取方法请参见**获取项目ID**。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
deployment_id	是	String	**参数解释：** 模型的部署ID，获取方法请参见**获取模型部署ID**。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

请求参数

表 3-59 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	**参数解释：** 用户Token。 用于获取操作API的权限。如Token认证中响应消息头中X-Subject-Token的值即为Token。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
Content-Type	是	String	**参数解释：** 发送的实体的MIME类型。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** application/json

表 3-60 请求 Body 参数

参数	是否必选	参数类型	描述
data	是	List<String>	**参数解释：** 待统计Token数的字符串。List长度必须为奇数。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

参数	是否必选	参数类型	描述
with_prompt	否	Boolean	**参数解释：** 是否仅统计输入字符的Token数。 **约束限制：** 不涉及 **取值范围：** ● true：仅统计输入字符串的Token数。 ● false：统计输入字符串和推理过程产生字符的总Token数。 **默认取值：** 不涉及

响应参数

表 3-61 响应 Body 参数

参数	参数类型	描述
tokens	List<String>	**参数解释：** 分解出的Token列表。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
token_number	Integer	**参数解释：** Token总数统计结果。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及

请求示例

```
{
  "data": [
```

```
      "你好，请介绍下西安。"
   ],
   "with_prompt": true
}
```

响应示例

```
{
   "tokens": [
      "你好",
      "，",
      "请",
      "介绍下",
      "西安",
      "。"
   ],
   "token_number": 6
}
```

状态码

请参见**状态码**。

错误码

请参见**错误码**。

4 附录

4.1 状态码

HTTP状态码为三位数，分成五个类别：1xx：信息响应；2xx：操作成功；3xx：重定向；4xx：客户端错误；5xx：服务器错误响应。

状态码如下所示。

状态码	编码	状态说明
100	Continue	继续请求。 这个临时响应用来通知客户端，它的部分请求已经被服务器接收，且仍未被拒绝。
101	Switching Protocols	切换协议。只能切换到更高级的协议。 例如，切换到HTTPS的新版本协议。
200	OK	服务器已成功处理了请求。
201	Created	创建类的请求完全成功。
202	Accepted	已经接受请求，但未处理完成。
203	Non-Authoritative Information	非授权信息，请求成功。
204	No Content	请求完全成功，同时HTTP响应不包含响应体。 在响应OPTIONS方法的HTTP请求时返回此状态码。
205	Reset Content	重置内容，服务器处理成功。
206	Partial Content	服务器成功处理了部分GET请求。
300	Multiple Choices	多种选择。请求的资源可包括多个位置，相应可返回一个资源特征与地址的列表用于用户终端（例如：浏览器）选择。

状态码	编码	状态说明
301	Moved Permanently	永久移动，请求的资源已被永久的移动到新的URI，返回信息会包括新的URI。
302	Found	资源被临时移动。
303	See Other	查看其他地址，使用GET和POST请求查看。
304	Not Modified	所请求的资源未修改，服务器返回此状态码时，不会返回任何资源。
305	Use Proxy	所请求的资源必须通过代理访问。
306	Unused	已经被废弃的HTTP状态码。
400	Bad Request	非法请求。 建议直接修改该请求，不要重试该请求。
401	Unauthorized	在客户端提供认证信息后，返回该状态码，表明服务端指出客户端所提供的认证信息不正确或非法。
402	Payment Required	保留请求。
403	Forbidden	请求被拒绝访问。 返回该状态码，表明请求能够到达服务端，且服务端能够理解用户请求，但是拒绝做更多的事情，因为该请求被设置为拒绝访问，建议直接修改该请求，不要重试该请求。
404	Not Found	所请求的资源不存在。 建议直接修改该请求，不要重试该请求。
405	Method Not Allowed	请求中带有该资源不支持的方法。 建议直接修改该请求，不要重试该请求。
406	Not Acceptable	服务器无法根据客户端请求的内容特性完成请求。
407	Proxy Authentication Required	请求要求代理的身份认证，与401类似，但请求者应当使用代理进行授权。
408	Request Timeout	服务器等待请求发生超时。 客户端可以随时再次提交该请求而无需进行任何更改。
409	Conflict	服务器在完成请求时发生冲突。 返回该状态码，表明客户端尝试创建的资源已经存在，或者由于冲突请求的更新操作不能被完成。
410	Gone	客户端请求的资源已经不存在。 返回该状态码，表明请求的资源已被永久删除。

状态码	编码	状态说明
411	Length Required	服务器无法处理客户端发送的不带Content-Length的请求信息。
412	Precondition Failed	未满足前提条件，服务器未满足请求者在请求中设置的其中一个前提条件。
413	Request Entity Too Large	由于请求的实体过大，服务器无法处理，因此拒绝请求。为防止客户端的连续请求，服务器可能会关闭连接。如果只是服务器暂时无法处理，则会包含一个Retry-After的响应信息。
414	Request URI Too Long	请求的URI过长（URI通常为网址），服务器无法处理。
415	Unsupported Media Type	服务器无法处理请求附带的媒体格式。
416	Requested Range Not Satisfiable	客户端请求的范围无效。
417	Expectation Failed	服务器无法满足Expect的请求头信息。
422	Unprocessable Entity	请求格式正确，但是由于含有语义错误，无法响应。
429	Too Many Requests	表明请求超出了客户端访问频率的限制或者服务端接收到多于它能处理的请求。建议客户端读取相应的Retry-After首部，然后等待该首部指出的时间后再重试。
500	Internal Server Error	表明服务端能被请求访问到，但是不能理解用户的请求。
501	Not Implemented	服务器不支持请求的功能，无法完成请求。
502	Bad Gateway	充当网关或代理的服务器，从远端服务器接收到了一个无效的请求。
503	Service Unavailable	被请求的服务无效。 建议直接修改该请求，不要重试该请求。
504	Gateway Timeout	请求在给定的时间内无法完成。客户端仅在为请求指定超时（Timeout）参数时会得到该响应。
505	HTTP Version Not Supported	服务器不支持请求的HTTPS协议的版本，无法完成处理。

4.2 错误码

当您调用API时，如果遇到“APIGW”开头的错误码，请参见**API网关错误码**进行处理。遇到“APIG”开头的错误码，请参考本文档进行处理。

表 4-1 错误码

模块	错误码	错误信息	说明	建议解决方法
模型推理	PANGU.0010	parameter illegal.	请求参数错误。	请参考《API文档》输入正确的请求参数，并重新调试API。
	PANGU.0011	Authentication failed.	认证失败。	认证鉴权失败，请参考《API文档》**认证鉴权**章节重新进行认证。
	PANGU.0012	The authentication information is missing.	缺少身份验证信息。	请检查调用API时是否有传入认证鉴权信息。
	PANGU.0031	Inner service exception.	服务内部异常。	请联系服务技术支持协助解决。
	PANGU.3254	The requested inference service does not exist.	资源不存在。	请检查调用API时projectId和deploymentId是否填写正确，推理服务状态是否可用。
	PANGU.3267	The number of service invoking requests exceeds the project limit.	用户调用过于频繁。	请降低请求频率。
	PANGU.3278	required api parameter is not present.	请求参数丢失。	请检查调用API时请求参数是否填写完整、是否有拼写错误、取值是否正确。
	PANGU.3318	The total length of the question should be between 1 and 4096.	Content长度不合法。	请参考《API文档》检查请求参数中输入的Content参数长度是否不在范围内，并重新调试API。
	PANGU.3320	The parameter [n] can only be 1 or 2 when calling non-streaming.	非流式调用推理服务传的参数只能是1或者2。	请使用正确的取值：1或者2。
	PANGU.3321	The parameter [n] can only be 1 when calling streaming.	流式调用推理服务n只能取1。	请使用正确的取值：1。

模块	错误码	错误信息	说明	建议解决方法
	PANGU.3342	Failed to invoke the inference service. please check the details field.	调用推理服务失败，请查看错误详情。	调用推理服务失败，请查看错误详情。
	IIT.0201	The input param is invalid!/The input param is invalid, please check your key!	请求参数不合法。	请检查请求参数是否填写正确。
	IIT.0202	Interval Server Error!	内部错误。	请联系服务技术支持协助解决。
	IIT.0203	The input param is invalid, the input data lens is less than the train data lens!	请求参数不合法，输入参数中的数据长度小于训练所用数据长度。	请确认请求body中特征名称、特征数量是否与训练数据中的特征一致。
	PREDICT.0102	Json format is wrong!或者其他与数据相关的特定报错信息	请求数据非JSON格式； 或者其他与数据相关的特定错误。	请将请求体设置为JSON格式； 或者根据数据相关的特定报错信息调整请求体。
	PREDICT.0201	The input param is invalid!/The input param is invalid, please check your key!	请求参数不合法。	请检查请求参数是否填写正确。
	PREDICT.0202	Interval Server Error!	内部错误。	请联系服务技术支持协助解决。
	PREDICT.0203	The input param is invalid, the input data lens is less than the train data lens!	请求参数不合法，输入参数中的数据长度小于训练所用数据长度。	请确认请求body中特征名称、特征数量是否与训练数据中的特征一致。
	APIG.0101	The API does not exist or has not been published in the environment.	访问的API不存在或尚未在环境中发布。	• 请检查API的URL是否拼写正确，例如，URL中是否缺少project_id。 • HTTP请求方法（POST，GET等）是否正确。

模块	错误码	错误信息	说明	建议解决方法
	APIG.0201	Backend timeout.	请求超时。	• 请检查原调用请求是否过于频繁，如果是并发过大，可以通过重试机制解决，在代码里检查返回值，碰到这个并发错误可以延时一小段时间（如2-5s）重试请求；也可以后端检查上一个请求结果，上一个请求返回之后再发送下一个请求，避免请求过于频繁。 • 请与技术支持确认，API是否已完成部署。
	APIG.0301	Incorrect IAM authentication information.	IAM身份验证信息不正确： • decrypt token fail：token解析失败。 • token expires：token过期。 • verify aksk signature fail：AK/SK认证失败。 • x-auth-token not found：未找到x-auth-token参数。	• token解析失败，请检查获取token的方法，请求体信息是否填写正确，token是否正确；检查获取token的环境与调用的环境是否一致。 • token超时（token expires），请重新获取token，使用不过期的token。 • 请检查AK/SK是否正确（AK对应的SK错误，不匹配；AK/SK中多填了空格）。 • AK/SK频繁出现鉴权出错，连续错误5次以上，被锁定5分钟（5分钟内，则一直认为其是异常的鉴权请求），5分钟后解锁重新认证。 • 检查账号权限，是否欠费，被冻结等。 • 检查调用API时，请求header参数X-Auth-Token是否拼写正确。

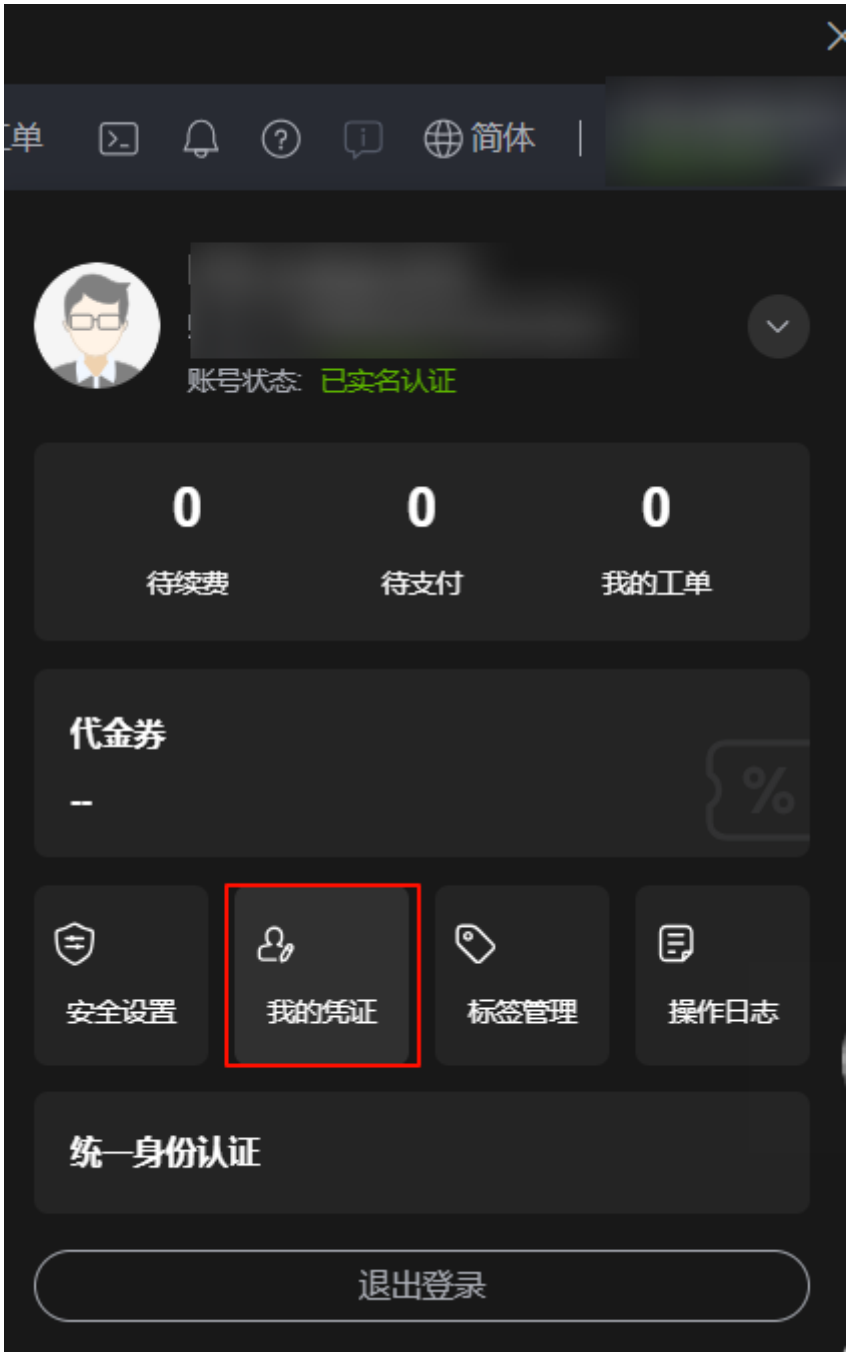
模块	错误码	错误信息	说明	建议解决方法
	APIG.0308	The throttling threshold has been reached: policy user over ratelimit,limit:XX,time:1 minute.	发送请求超过了服务的默认配置限流。	• 通过重试机制，在代码里检查返回值，碰到并发错误可以延时一小段时间（如2-5s）重试请求。 • 后端检查上一个请求结果，上一个请求返回之后再发送下一个请求，避免请求过于频繁。

4.3 获取项目 ID

从控制台获取项目 ID

1. 登录**管理控制台**。
2. 在页面右上角的用户名的下拉列表中选择“我的凭证”。

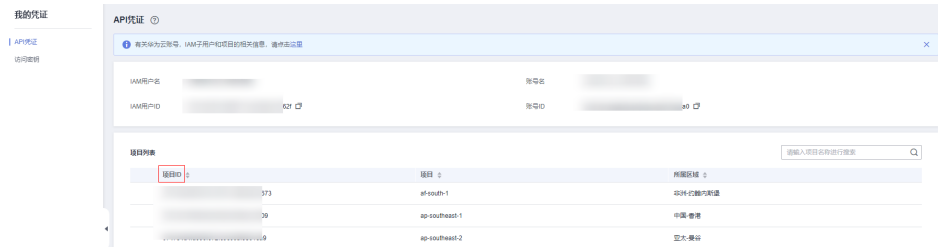
图 4-1 我的凭证



3. 在“我的凭证”页面，获取项目ID（project_id），以及账号名、账号ID、IAM用户名和IAM用户ID。

   在调用盘古API时，获取的项目id需要与盘古服务部署区域一致，例如盘古大模型当前部署在“中国-香港”区域，需要获取与香港区域的对应的项目id。

图 4-2 查看项目 ID



多项目时，展开“所属区域”，从“项目ID”列获取子项目ID。

调用 API 获取项目 ID

项目ID还可通过调用**查询指定条件下的项目信息**API获取。

获取项目ID的接口为“GET https://{endpoint}/v3/projects”，其中{endpoint}为IAM的终端节点，可以从**地区和终端节点**获取。接口的认证鉴权请参见**认证鉴权**。

响应示例如下，例如，对话机器人服务部署的区域为“ap-southeast-1”，响应消息体中查找“name”为“ap-southeast-1”，其中projects下的“id”即为项目ID。

```
{
    "projects": [
        {
            "domain_id": "65382450e8f64ac0870cd180d14e684b",
            "is_domain": false,
            "parent_id": "65382450e8f64ac0870cd180d14e684b",
            "name": "project_name",
            "description": "",
            "links": {
                "next": null,
                "previous": null,
                "self": "https://www.example.com/v3/projects/a4a5d4098fb4474fa22cd05f897d6b99"
            },
            "id": "a4a5d4098fb4474fa22cd05f897d6b99",
            "enabled": true
        }
    ],
    "links": {
        "next": null,
        "previous": null,
        "self": "https://www.example.com/v3/projects"
    }
}
```

4.4 获取模型部署 ID

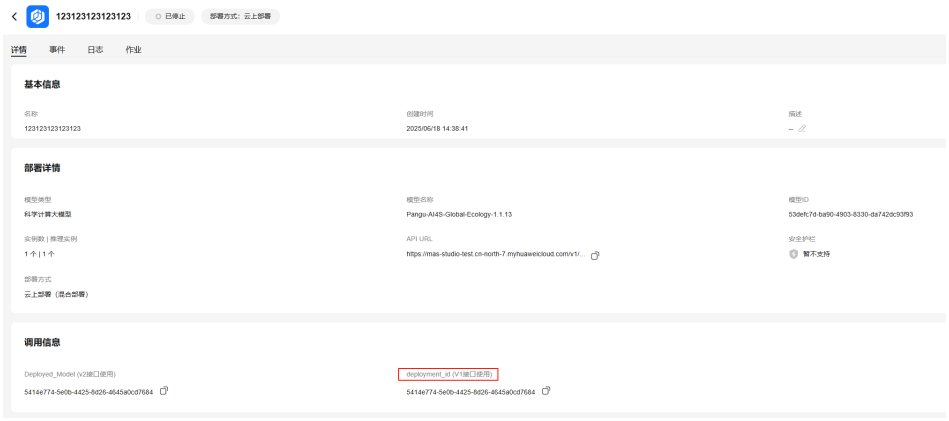
模型部署ID获取步骤如下：

步骤1 登录ModelArts Studio大模型开发平台。

步骤2 获取模型部署ID。

- 若调用部署后的模型，可在左侧导航栏中选择“模型开发 > 模型部署”，在“我的服务”页签，模型部署列表单击模型名称，在“详情”页签中，可获取模型的部署ID。

图 4-3 部署后的模型调用路径



- 若调用预置模型，可在左侧导航栏中选择“模型开发 > 模型部署”，在“预置服务”页签，模型列表单击“调用路径”，获取该模型的部署ID。

图 4-4 预置模型的部署 ID

----结束